CALIFORNIA ASSOCIATION FOR INSTITUTIONAL RESEARCH

NOVEMBER 2019

THE CAIR REPORT

Institutional Research for the Public Good

Volume 1 • Issue 1

In This Issue:

Letter from the Editorial Board

Welcome to the inaugural volume of *The CAIR Report*, an annual online publication that showcases presentations from CAIR's annual conference in written form. Each volume of *The CAIR Report* will reflect the theme of the previous year's conference. The first volume represents the 2018 CAIR Conference theme of "Institutional Research for Public Good." Articles in this volume of *The CAIR Report* feature a variety of topics IR professionals may take on (i.e. assessment, faculty evaluations and academic support decisions) in support of the mission of their institutions.

The CAIR Report offers the opportunity for institutional researchers to publish work that is particularly relevant for our community. Papers submitted for publication were reviewed by an editorial board of former CAIR presidents who have a wealth of experience in the field. Authors present statistical techniques, strategies for working interdepartmentally, and how their work is used for decision-making on campus.

In this issue of *The CAIR Report*:

- John Kim, Ludmila Praslova, and Philip Newlin at Vanguard University of Southern California explore the myths surrounding student ratings of professors, and outline how an IR office can support teaching effectiveness and student learning measures in "Reliability and Validity of Instructor Course Evaluations: Exploring the Myths",
- Jose de los Reyes at San Francisco Art Institute shares his approach for IR collaboration with faculty and other departments at a small arts education institution to determine if diversity-related outcomes are in place in the academic curriculum in "Assessing Diversity-Related Outcomes using Course Syllabi", and
- Heidi Carty and Galina Belokurova at University of California, San Diego investigate how predictive modeling can be used to identify academically at-risk student athletes in "Using Predictive Analytics to Identify 'At Risk' Student Athletes".

The CAIR Report editorial board is committed to representing the research and experiences of new and veteran institutional research professionals from a diverse array of institutions. We hope this inaugural volume of *The CAIR Report* inspires creative thinking in Institutional Research on your campus and the sharing of your institution's resultant work at CAIR conferences in the future.

Sincerely,

The CAIR Report Editorial Board:
Editor-in-Chief: Brianna B. Moore-Trieu, PhD. CAIR President, 2018
Associate Editors:
Robert Daly CAIR President, 1987
Jessica Luedtke, CAIR President, 2019
Alice van Ommeren, Ed D. CAIR President, 2013
Kristina Powers, Ph D. CAIR President, 2016
Juan Ramirez, Ph D. CAIR President, 2017

Special thanks to Deborah Lee and Judith Teruya for the aesthetic design of The CAIR Report.

Reliability and Validity of Instructor Course Evaluations: Exploring the Myths

John Kim, Philip Newlin, Ludmila Praslova
Vanguard University of Southern California

Abstract

Course evaluations are widely implemented in higher education, but there are numerous criticisms regarding their reliability and validity. With questions about instructional effectiveness and degree completion for a changing undergraduate demographic, it is essential for both instructors and administrators to have reliable feedback about instructional quality. The current study examined reliability and validity of course evaluations and explored some of the myths surrounding student ratings of teaching effectiveness using four semesters of course evaluation results from a medium-sized faith-based university in Southern California. Results showed high reliability and acceptable evidence of construct validity for the course evaluations. The findings indicate that student ratings of instruction can provide useful formative insight for developing teaching skills. However, results also showed that individual course evaluation scores could be greatly affected by extraneous variables such as course discipline and faculty ethnicity, especially when the number of respondents is small. Therefore, careful interpretation of course evaluation results and understating the influence of bias on ratings of instruction are essential to evaluate and improve faculty teaching effectiveness. We suggest using the average course evaluation results with a sufficiently large number of responses from three or four consecutive terms as the larger sample can help control for extraneous variables and the longitudinal data can illuminate changes in teaching effectiveness over time.

INTRODUCTION

In most higher education institutions, instructor course evaluations are one of the most widely used measures of teaching effectiveness and are a heavily weighted criterion for faculty promotion (Algozzine et al., 2004; Bassett, Cleveland, Acorn, Nix, & Snyder, 2017; Cashin, 1988; Hobson & Talbot, 2001). However, the usefulness and validity of student ratings of instruction have frequently been challenged. Several authors consider instructor course evaluations (or student evaluations of teaching) unreliable and invalid due to several sources of student bias, such as gender, race, academic discipline, student motivation, and expected grade (Bassett et al., 2017; Boring, Ottoboni, & Stark, 2016; Flaherty, 2016). This leads to the argument against the use of course evaluations for personnel decision making (Boring et al., 2016; Falkoff, 2018; Flaherty, 2016; Stark & Freishtat, 2014). Proponents of course evaluations maintain that misgivings about student ratings of instructors are not supported by research and that instructor course evaluations are more valid than any other measures of teaching effectiveness (Benton & Cashin, 2012; Cohen, 1981; Marsh, 2007).

One of the most frequent criticisms of course evaluations is potential gender bias with evidence to suggest that students rate male professors more favorably than they do female professors and with gender bias more prevalent among male students (Benton & Cashin, 2012; Falkoff, 2018; Flaherty, 2016; MacNell, Driscoll, & Hunt, 2015). Along with gender bias, race (or ethnicity) has been investigated as a source of bias in several studies. A cluster analysis of data obtained from www.ratemyprofessors.com revealed that students rated racial minority faculty less favorably, especially African Americans and Asians (Reid, 2010). Moreover, students tended to rate African American (Harlow, 2003; Hendrix, 1997, 1998; Ho, Thomsen, & Sidanius, 2009) or African American male faculty (Reid, 2010) less favorably than they did White faculty. Feldman, Centra (2009), and many researchers (as cited in Benton & Cashin, 2012) also found that math and science courses or any disciplines requiring quantitative reasoning skills received lower ratings than other disciplines, especially arts and humanities.

Therefore, this study aimed to 1) examine the reliability and validity of instructor course evaluations, 2) investigate some of the stereotypical assumptions regarding course evaluations, and 3) identify the subscales of the course evaluation survey through factor analysis and item analysis of each subscale using multi-year course evaluation data of a faith-based university in Southern California.

METHODS

Participants and Courses

A total of 28,181 student responses to 1,364 traditional undergraduate course evaluations were collected over four semesters (2016 fall through 2018 spring) from a private, not-for-profit, religiously affiliated university. Only regular traditional undergraduate courses were used, and all non-regular courses (internship, practicum, independent study, etc.) were excluded from data analysis. Of the 1,364 courses, 694 (51%) were lower division, and 670 (49%) were upper division; 133 (9.8%) were Business, 110 (8.1%) were Communication, 197 (14.4%) were Fine

Arts, 130 (9.5%) were Language and Education, 189 (13.9%) were Religion, 307 (22.5%) were Social Science, and 298 (21.8%) were STEM (Chemistry, Biology, Mathematics) and Kinesiology courses.

A total of 189 instructors were evaluated; 70 (37%) were female, 98 (52%) were male, and 21 (11%) were unknown. Details on the faculty race/ethnicity, type, and degree are provided in Table 1.

Table 1
Race/Ethnicity, Type, and Degree of Faculty

Characteristics	*n*	%
Race/Ethnicity		
White	114	60.3
Asian	15	7.9
Hispanic	12	6.3
African American	6	3.2
Others (two or more races, non-resident alien, unknown)	42	22.2
Faculty Type		
Adjunct	108	57.1
Tenured	42	22.2
Tenure-Track	20	10.6
Term-Contract	19	10.1
Faculty Degree		
Doctoral	83	43.9
Terminal (highest degree in a field)	10	5.3
Master	54	28.6
Bachelor	18	9.5
Unknown	24	12.7

Survey Questionnaire

Course evaluation survey questions were developed by a group of faculty, including the Associate Provost, as part of adopting a new online course evaluation system, EvaluationKIT (www.evaluationkit.com/). The survey includes 18 items: 10 questions asking students' opinions about the instructor's class presentation, syllabus preparation, helpfulness of assignments, instructor's availability to help, and faith–class material integration expected by the institution in alignment with its religions affiliation using a Likert-style responding format with five alternatives (see Table 2). In addition, eight open-ended or multiple-choice questions with nominal options ask students' opinions about the effectiveness of teaching aids, class difficulty, hours of self-study, final degree goal, and suggestions. This study analyzed only the first 10 Likert-scale items.

Table 2
Instructor Course Evaluation Survey Questions

Item #	*Question Content*
1	Explaining the course requirements
2	Preparation for each class session
3	Effective class time management
4	Exhibiting Christian worldview
5	Integration of faith with course content
6	Responsiveness to questions
7	Availability to help outside of the classroom
8	Grading & returning assignments in a reasonable amount of time
9	Following course syllabus for the course content and the pace
10	Helpfulness of the assigned class activities for learning

Likert Scale: 1=strongly disagree, 2=disagree, 3=not sure, 4=agree, 5=strongly agree

Data Analyses

First, the key psychometric properties of the course evaluation survey, such as reliability and validity, were examined. Reliability refers to the accuracy and consistency of a measure, which is a prerequisite for validity (Mitchell & Jolley, 2010). Internal consistency using Cronbach's alpha is the most general method of estimating reliability. Cronbach's alpha of .80 or more is typically considered to indicate high reliability, suggesting that 20% or less of the variability in scale scores is due to measurement error. Scale validity can be defined as "the agreement between a test score or measure and the quality it is believed to measure" (Kaplan & Saccuzo, 2001, p. 132). In other words, it is the degree to which the scale measures what it is supposed to measure. Among the different types of validity evidence (content, criterion, and construct), construct validity can be defined as an "appropriateness of inference from test scores to latent construct" (Embretson, 2008). The latent construct that an instructor course evaluation seeks to measure should obviously be the "True Teaching Effectiveness" of an instructor (denoted by θ henceforward). Then, θ serves as the construct-related evidence for validity. However, there are several possible operational definitions of θ. In this study, an average of all the course evaluation scores of an instructor for four semesters, from 2016 fall to 2018 spring, was used as the operational definition of θ (denoted by $\hat{\theta}$ henceforward). The $\hat{\theta}$ was chosen based on the assumption that a more effective instructor should, on average, receive higher evaluation scores and that a sufficiently large dataset could effectively control for random and extraneous variables. Additionally, only evaluation results from courses with at least 10 students were used to minimize the confounding effect of possible aberrant individual response patterns.

Criterion-related validity shows how well a measure predicts or corresponds with a well-defined criterion measure (Kaplan & Saccuzo, 2001). For example, college GPA can serve as criterion-related evidence of SAT score accuracy. In many studies, student learning (or achievement) has been identified as the best criterion-related evidence of teaching effectiveness (Benton & Cashin, 2010; Bassett et al., 2017; Cohen, 1981; Marsh, 2007). Theoretically, students taught by more effective instructors should outperform students taught by less effective instructors when other conditions are same; therefore, final course grade can serve as one of the possible operational definitions of student learning (Benton et al., 2011; Cohen, 1981, 1987; Feldman, 1989; Marsh, 2007). In this study, criterion-related validity of the instructor course evaluation was checked using both individual course final grade and a multi-year average of course final grade of each instructor. The effect size of the correlation coefficient in the current study was determined based on Cohen's (1988) guideline: small=0.1, medium=0.3, large=0.5.

Second, univariate regression analyses were conducted to test some of the stereotypical assumptions about course evaluations. In the regression model, the effect of each of the following eight factors was tested for its significance after controlling for all the other factors:

- Gender: male and female
- Ethnicity: African American, Asian, Hispanic, and White (using dummy codes with White as a reference category)
- Academic Discipline: Business, Humanities, Fine Arts, Social Science, Natural Science, and Religion (using dummy codes with Religion as a reference category)
- Class Size: ranging 2 to 98
- Course Division: lower and upper
- Course Final Grade: ranging from 1.5 to 4.0
- Faculty Type: Tenured, Tenure-track, Term-contract, and Adjunct (using dummy codes with Tenured as a reference category)
- Faculty Degree: Doctoral, Terminal, Master's, and Bachelor's (using dummy codes with Doctoral as a reference category)

Regression modeling was replicated for four semesters to confirm the external validity of the findings. Any factors that were significant for at least three semesters were considered strong predictors of course evaluation results.

Third, exploratory factor analyses using the principal axis factoring method was conducted to determine the number of factors in the course evaluation survey. The factor analysis utilized all individual responses (N=28,181) on the 10 survey questions for 1,364 courses over four semesters. Eigenvalue (≥1), cumulative percent of common variance extracted (75% to 85%), scree plot, and interpretability of the factors were considered as criteria for choosing the optimal number of factors. For a cumulative proportion of common variance, Gorsuch (1983) suggests 75% to 85% as appropriate cut-off points to decide the number of factors unless a significant amount of the variance can be accounted for by an additional factor. Also, since the underlying latent common factors were intercorrelated, an oblique rotation method was chosen (Promax with Kaiser Normalization) to identify the subscales. Although the normalized varimax method of orthogonal rotation is popular for exploratory factor analysis, the orthogonality assumption of common factors is an artifact of the method because the assumption is rarely met in real-world cases (Mulaik, 2010). Separate factor analyses were conducted using the same procedure for four semesters to cross-validate the factor structure and subscale items.

RESULTS

Reliability and Validity

The course evaluation questionnaire showed very high internal consistency with Cronbach's alphas ranging from .93 to .94 for the four consecutive semesters, 2016 fall through 2018 spring (see Table 3). For the construct-related validity, an estimator of true teaching effectiveness ($\hat{\theta}$) was calculated for each instructor by averaging their course evaluation scores for four semesters. Only courses with at least 10 respondents were included for the calculation, yielding $\hat{\theta}$s of 42 instructors for a total of 440 courses. The correlation between individual course evaluation scores and $\hat{\theta}$ was significant (p<.01) with a large effect size (r>.50) for all four semesters (r=.62, r=.75, r=.70, and r=.77 for 2016 fall, 2017 spring, 2017 fall, and

2018 spring, respectively). These results indicate very acceptable evidence of construct validity of the course evaluation scores.

Table 3
Internal Consistency Reliability

Term	*α*	*N of Items*	*N of Courses*	*N of Respondents*
2016 F	.941	10	328	5535
2017 S	.926	10	330	5076
2017 F	.941	10	360	5910
2018 S	.939	10	346	5291

Research (Benton et al., 2011; Cohen, 1981, 1987; Feldman, 1989; Marsh, 2007) supports the theory that students taught by more effective instructors will outperform students taught by less effective instructors. This theory is also supported by the current study with a significant correlation between $\hat{\theta}$ and the average of all course final grades over four semesters for each instructor (r=.34, p<.05) with a medium effect size (r>0.3); however, very low or no correlation was observed between individual course evaluation scores and the average course final grade (r=.13, r=.16, r=.05, and r=.33 for 2016 fall through 2018 spring, respectively). Similarly, very low correlations were observed between individual course evaluation scores and individual course final grades, (r=.11, r=.09, r=.07, and r=.12 for 2016 fall through 2018 spring, respectively). Results of the current study do not support the criterion-related validity of course evaluation. It was speculated that individual course evaluation scores, unlike the average course evaluation scores, could easily be affected by random and extraneous variables. Therefore, it is important to know which extraneous variables significantly affect the internal validity of course evaluation.

Regression Analysis

Regression modeling showed that the academic discipline of the course, faculty ethnicity, and faculty degree were strong predictors of course evaluation responses after controlling for all other factors. Religion courses showed significantly higher course evaluation scores than Business and Natural Science courses for all four semesters and higher scores than Humanities courses (communication, English, and Spanish) for three semesters (see Table 4). White faculty received significantly higher ratings than Asian and African American faculty but were not rated significantly differently from Hispanic faculty for three semesters. These findings are consistent with the study by Reid (2010), which states that Asian and African American faculty are rated less favorably when compared to White faculty. It was speculated that the findings of this study may also be attributed to the ethnic composition of the student population, which is 3% Asian, 5% African-American, 39% White, and 42% Hispanic as of 2017.

Faculty with doctoral degrees showed significantly higher scores than faculty with terminal masters or professional degrees for three semesters, suggesting that faculty degree is a significant predictor of evaluation scores. All the other factors, such as gender, faculty type, course division, class size, and course final grade, showed no or inconsistent significance (replicated for two or less semesters) in this study. No significant difference in course evaluation scores was found between male and female faculty for three semesters, and female faculty were rated significantly higher than male faculty in the 2018 spring semester, which contrasts the claims suggesting that course evaluations favor male instructors. The predominantly female student gender composition could in part explain these findings; future research should examine respondent characteristics in addition to instructor characteristics.

Factor Analysis

Exploratory factor analysis verified the three-factor structure of the course evaluation survey. Two or three factors seemed to be feasible options: Eigenvalues (1.17 for two-factor model, 0.671 for three-factor model), cumulative percent of common variance extracted (78.0% for two factor, 84.7% for three factor), and scree plot (see Figure 1) where both numbers two and three on the x-axis could be considered good leveling-off points. However, since three subscales provided a better interpretability of the survey structure, a three-factor model was ultimately chosen. The items of each factor (subscale) and their factor loadings are provided in Table 5. The three subscales of the survey were identified based on the three factors obtained from the exploratory factor analysis. As shown in Table 5, the instruction-related scale comprised five items (1, 2, 3, 6, and 7), the faith-related scale comprised two items (4 and 5), and the assignment-related comprised three items (8,9, and 10). The three-factor model clearly showed a simple structure with each variable loading on one or as few factors as possible with at least one zero loading (Thurston, 1935). Item 7 (Instructor’s availability to help outside of the classroom) loaded on factor 1

Table 4

Regression Coefficients of the Course Evaluation-Related Factors and the Significance

Factors	2016FA		2017SP		2017FA		2018SP	
	β	*p-value*	β	*p-value*	β	*p-value*	β	*p-value*
Class Size	−0.11	0.111	−0.07	0.313	−0.15	0.016	−0.06	0.268
Course Division (Lower vs. Upper)	−0.14	0.027	−0.04	0.583	−0.22	0.001	−0.11	0.060
Course Final Grade	0.18	0.003	0.09	0.161	0.06	0.316	0.17	0.002
Discipline (Business vs. Religion)	−0.24	0.002	−0.25	0.001	−0.16	0.032	−0.14	0.042
Discipline (Humanities vs. Religion)	−0.23	0.013	−0.29	0.001	−0.13	0.124	−0.20	0.012
Discipline (Fine Arts vs. Religion)	−0.13	0.138	−0.13	0.124	−0.19	0.028	0.00	0.965
Discipline (Social Science vs. Religion)	−0.09	0.318	−0.14	0.098	−0.12	0.134	−0.12	0.100
Discipline (Natural Science vs. Religion)	−0.19	0.028	−0.29	0.001	−0.42	0.000	−0.34	0.000
Faculty Degree (Bachelor or below vs. Doctoral)	−0.10	0.142	−0.06	0.364	−0.04	0.585	−0.10	0.086
Faculty Degree (Master vs. Doctoral)	−0.04	0.637	−0.07	0.355	−0.16	0.025	−0.19	0.006
Faculty Degree (Terminal vs. Doctoral)	−0.16	0.014	0.03	0.696	−0.13	0.038	−0.26	0.000
Faculty Ethnicity (Asian vs. White)	−0.22	0.000	−0.12	0.041	−0.10	0.079	−0.14	0.009
Faculty Ethnicity (African American vs. White)	−0.15	0.012	−0.03	0.566	−0.29	0.000	−0.27	0.000
Faculty Ethnicity (Hispanic vs. White)	−0.07	0.239	−0.24	0.000	−0.01	0.872	−0.10	0.073
Faculty Ethnicity (Two or more Races vs. White)	−0.05	0.444	0.02	0.695	−0.04	0.445	−0.17	0.002
Faculty Gender (Female vs. Male)	−0.01	0.882	−0.08	0.178	−0.02	0.698	−0.21	0.000
Faculty Type (Adjunct vs. Tenured)	−0.03	0.723	0.07	0.356	0.07	0.392	0.12	0.111
Faculty Type (Term Contract vs. Tenured)	0.00	0.995	0.20	0.005	−0.02	0.716	0.00	0.983
Faculty Type (Tenure Track vs. Tenured)	−0.10	0.147	−0.06	0.339	−0.03	0.607	−0.06	0.330

Note: β=standardized coefficient

Figure 1. Scree Plot of the Exploratory Factor Analysis.

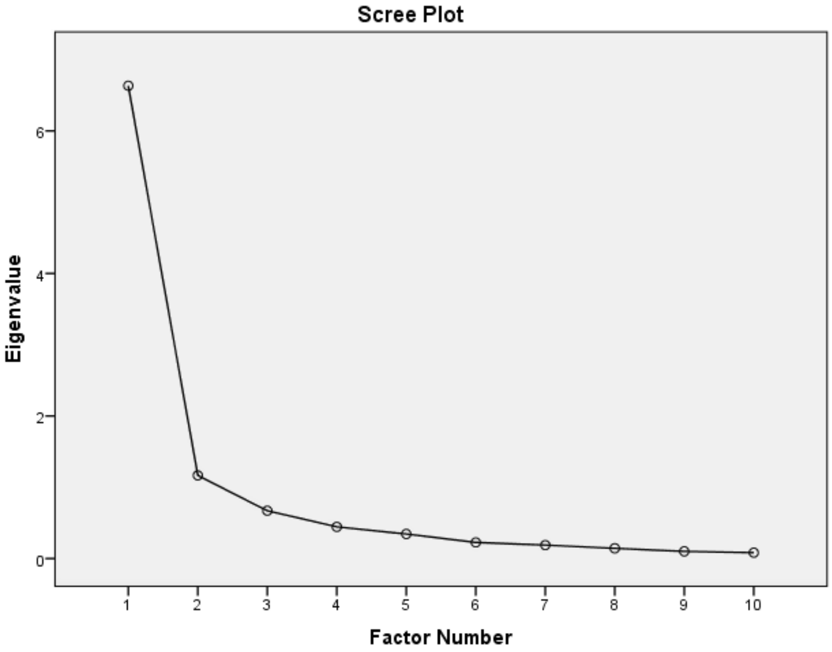


Table 5
Items of Each Subscale and Their Factor Pattern

Subscale	Item #	Factor 1	Factor 2	Factor 3	Item–Scale Corr.
Instruction-related (α=.94)	3	1.000	−0.084	−0.028	0.867
	2	0.928	−0.065	0.033	0.863
	1	0.773	0.011	0.16	0.867
	6	0.728	0.236	−0.043	0.824
	7	0.445	0.331	0.084	0.714
Faith-related (α=.92)	4	−0.025	0.988	−0.024	0.852
	5	−0.013	0.925	0.039	0.852
Assignment-related (α=.86)	9	0.108	−0.097	0.925	0.664
	8	−0.074	0.097	0.728	0.826
	10	0.398	0.066	0.467	0.738

Note: by principal axis factoring and the Promax rotation

(Instruction-related) and factor 2 (faith-related) with the loadings of .445 and .331, respectively. It was speculated that instructor's availability to help students outside of the classroom could be interpreted by students as instructor's faithfulness. Similarly, item 10 (Helpfulness of the assigned class activities for learning) was cross-loaded on factors 1 and 3 (assignment-related) with the loadings of .398 and .467, respectively. This may suggest that the words "assigned class activities" should be modified to be more specific since they could be interpreted as either a part of instruction or a regular assignment. The item–scale correlations were also provided in Table 5 along with the Cronbach's alpha (α) within each subscale. All items showed strong item–scale correlations ranging from .664 to .867 with very high internal consistency for each subscale (.94, .92, and .86 for instructional-related, faith-related, and assignment-related scales, respectively). Finally, the three-factor structure with the corresponding items was clearly replicated across four semesters, thus verifying the external validity of the three subscales.

DISCUSSION

This study aimed to examine the reliability and validity of instructor course evaluations and to investigate several sources of bias affecting student ratings of instruction. Results showed very high reliability and strong construct-related validity of the course evaluation survey; however, criterion-related validity was not acceptable. It was an interesting finding that the average course evaluation score of an instructor was significantly correlated with the average course final grade of the instructor while individual course evaluation score was minimally or not at all correlated with the average course final grade, consistent across all four semesters. The result contrasts with the popular notion that course evaluations are simply a reflection of the student's anticipated grade. These findings indicate that student ratings of instruction can still provide insight that improves teaching ability; however, it is necessary to examine an instructor's full body of work rather than scrutinizing a discrete classroom experience.

Results also show that individual course evaluation scores are clearly affected by random and extraneous variables other than teaching effectiveness, especially when the number of respondents is small. Our pilot study showed that datasets including course evaluations with fewer than 10 respondents were misleading. Therefore, it is important to consider course evaluation results with a sufficiently large number of responses. Additionally, we would suggest that the average course evaluation results of multiple years should be used as evidence of the teaching effectiveness of each faculty member as the average result is much less likely to be affected by extraneous variables. Average course evaluation results from the recent three or four consecutive terms seems ideal as the sample is sufficiently large to control for extraneous effects but also accommodates changes in teaching effectiveness over time.

Limitations of the current study were that the results were found at one institution and that the respondent characteristics (e.g., student gender and ethnicity) were unavailable. Future research should explore the external validity of the findings in the current study at different types of institutions and should examine the interaction between respondent and instructor characteristics to more accurately assess potential sources of bias in student ratings. Finally, additional objective criteria other than average final course grade will be necessary to confirm the validity of the course evaluation measure. Additional studies utilizing longitudinal design and objective measures of performance may shed further light on the relationship between student evaluation of teaching, teaching approaches that support the immediately rewarding course experience, and teaching approaches that facilitate deep and long-term learning. This would, in turn, inform institutional approaches to faculty development. Offices of institutional research could make a major contribution to advancing teaching and learning in higher education by collecting and analyzing longitudinal data on course evaluation and grades as well as collaborating with institutional assessment of student learning efforts and adding the data from direct assessment of student learning, including learning in subsequent courses, to statistical models for understanding the relationship of teaching and student achievement.

REFERENCES

Algozzine, B., Beattie, J., Bray, M., Flowers, F., Gretes, J., Howley, L., Mohanty, G., & Spooner, F. (2004). Student evaluation of college teaching: A practice in search of Principles. *College Teaching, 52(4)*, 134–141.

Bassett, B., Cleveland, A., Acorn, D., Mix, M., & Snyder, T. (2017). Are they paying attention? Students' lack of motivation and attention potentially threaten the utility of course evaluations. *Assessment & Evaluation in Higher Education, 42(3)*, 431–442.

Benton, S. L., & Cashin, W. E. (2010). *Student ratings of teaching: A summary of research and literature* (IDEA Paper No. 50). Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Boring, A., Ottoboni, K., & Stark, P. B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research, 1*.

Braga, M., Paccagnella, M., & Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review, 41*, 71–88. 10.1016/j.econedurev.2014.04.00

Carrell, S. E., & West, J. E. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy, 118*, 409–432. 10.1086/653808

Cashin, W. E. (1988). *Student ratings of teaching: A summary of the research* (IDEA Paper No. 20). Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Centra, J. A. (2009). *Differences in responses to the student instructional report: Is it bias?* Princeton, NJ: Educational Testing Service.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multi-section validity studies. *Review of Educational Research, 51*, 281–309.

Embretson, S. E. (2008). *Psychometric theory* [Class notes]. Atlanta, GA: Georgia Institute of Technology, PSYCH7303.

Falkoff, M. (2018, April 26). Why we must stop relying on student ratings of teaching. *The Chronicle of Higher Education.* Retrieved June 26, 2018, from *https://www.chronicle.com/article/Why-We-Must-Stop-Relying-on/243213*

Flaherty, C. (2016, January 11). Bias against female instructors. *The Inside Higher Ed.* Retrieved January 12, 2016, from https://www.insidehighered.com/news/2016/01/11/new-analysis-offers-more-evidence-against-student-evaluations-teaching.

Gorsuch, R. L. (1983). *Factor Analysis* (2nd ed.). Hillsdale, NJ: Lawrence.

Halpern, D. F., & Hakel, M. D. (2003). Applying the science of learning to the university and beyond: Teaching for long-term retention and transfer. *Change*, July/August, 2–13.

Harlow, R. (2003). "Race doesn't matter, but…": The effect of race on professors' experiences and emotion management in the undergraduate college classroom. *Social Psychology Quarterly, 66*, 348–363.

Hendrix, K. G. (1997). Student perceptions of verbal and nonverbal cues leading to images of black and white professor credibility. *Howard Journal of Communications,* 8, 251–273.

Hendrix, K. G. (1998). Student perceptions of the influence of race on professor credibility. *Journal of Black Studies, 28*, 738–763.

Ho, A. K., Thomsen, L., & Sidanius, J. (2009). Perceived academic competence and overall job evaluations: Students' evaluations of African American and European American professors. *Journal of Applied Social Psychology, 39*, 389–406.

Hobson, S. M. & Talbot, D. M. (2001). Understanding student evaluations. *College Teaching, 49,* 26–31.

Kaplan, R. M., & Saccuzzo, D. P. (2001). *Psychological testing: Principles, applications, and issues* (5th ed.). Belmont, CA: Wadsworth.

MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a name? Exposing gender bias in student ratings of teaching. *Innovative Higher Education, 40*, 291–303.

Marsh, H. W. (2007). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (eds.) *Scholarship of teaching and learning in higher education*: *An evidence-based perspective*, 319–384. New York: Springer.

Mitchell, M., & Jolley, J. (2010). *Research design explained* (7th ed.). Bellmont, CA: Wadsworth.

Mulaik, S. A. (2010). *Foundations of factor analysis* (2nd ed.). Boca Raton, FL: CRC Press.

Praslova, L. (2010). Adaptation of Kirkpatrick's four level model of training criteria to assessment of learning outcomes and program evaluation in Higher Education. *Educational Assessment, Evaluation and Accountability,* 22, 215–125, Springer Netherlands. DOI: 10.1007/s11092-010-9098-7

Reid, L. D. (2010). The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors.com. *Journal of Diversity in Higher Education, 3*, 137–152.

Stark, P. B. & Freishtat, R. (2014). *An evaluation of course evaluations.* Berkley, CA: Center for Teaching and Learning

Thurston, L. L. (1935). *The vectors of the mind.* Chicago: University of Chicago Press.

Assessing Diversity-Related Outcomes with Course Syllabi

Jose de los Reyes
San Francisco Art Institute

Abstract

At many colleges, diversity is mentioned as an institutional value or associated with learning outcomes. How can an institution determine if diversity-related outcomes are in place in the academic curriculum? A rich source of data for this assessment is course syllabi. This paper presents a self-study led by the Institutional Research Office at the San Francisco Art Institute evaluating the presence of diversity-related outcomes in the curriculum through a review of course syllabi. In the self-study, institutional research coordinated the collection of syllabi and provides the faculty with expertise in quantitative techniques for 1) improving assessment process, i.e., norming rater scores in the presence of rater bias, and 2) interpreting the results of the assessment. Results of the self-study along with lessons learned about institutional research interfacing with assessment are discussed. Institutional researchers initiating work within assessment at their own institutions may find this framework useful for their approach.

INTRODUCTION

The San Francisco Art Institute (SFAI) conducts an annual program assessment in the form of a self-study. In 2015, the Provost requested that the topic of the self-study be diversity-related outcomes in the academic curriculum. The project was to be undertaken by the Program Assessment Committee of the Faculty Senate, which typically comprises faculty members, the Dean of Academic Affairs, and several Academic Affairs staff members. The Program Assessment Committee appointed the Institutional Research Associate as the lead researcher for the self-study. The first task was to review the institutional definition of diversity. Diversity is defined and explored at SFAI in the school's Diversity Statement published in 2012:

> A rigorous artistic and intellectual community is enriched by diversity and inclusion. We promote artistic and intellectual freedom by fostering environments that value. . . and provide all community members with a respectful and challenging space in which to address divergent opinions and ideas.
>
> By “diversity,” we mean that our community prospectively embraces. . . the many forms of composite subjectivity and life experience that span these [characteristic] differences. Promoting such a broadly inclusive understanding of diversity requires ongoing education and effort to ensure support understanding and awareness from all community members. In this, SFAI strives to move beyond the reactive methodologies of affirmative action, even as we proactively practice equal opportunity in hiring and admissions.
>
> SFAI seeks to be a vanguard institution with regard to how we address and integrate notions of diversity. The Institute continues to develop connections and mutually beneficial relationships between the school’s immediate community and local and global publics in the belief that a multiplicity of voices has helped to make SFAI the influential and inspiring institution that it is today.

Initial challenges for the study were to develop a tool that incorporates the institution's definition of diversity to assess diversity outcomes in the curriculum and to use pre-existing data for assessment.

The Program Assessment Committee[1] developed a diversity rubric for syllabi that allowed evaluators to score components of a syllabus against diversity outcomes. The rubric was based on both SFAI's Bachelor of Fine Arts (BFA) assessment rubric and the *Self-Assessment Rubric for the Institutionalization of Diversity, Equity and Inclusion in Higher Education* by the New England Resource Center for Higher Education's Multicultural Affairs Think Tank.

Next, institutional research inventoried various offices and departments for course assessment tools assessing diversity-related outcomes. For example, it was discovered that some departments had begun to introduce diversity-specific questions in course evaluations. Some had also launched assessment tools to directly address diversity, such as a Diversity Survey utilized in a foundational art history course. After

[1] The Program Assessment Committee was comprised of: Dr. Nicole Archer, Assistant Professor; Tamara Loewenstein, BA & BFA Department Manager; Jennifer Rissler, Interim Dean and Vice President of Academic Affairs; Mark Van Proyen, Associate Professor; Lindsey White, Assistant Professor; and Jose de los Reyes, Institutional Research & Academic Planning Associate.

reviewing all available assessment tools, the committee decided to develop a rubric to establish a standard and sustainable method for assessing diversity in the curriculum and called it the Diversity Rubric for Course Syllabi.

The Diversity Rubric for Course Syllabi (see Appendix) has six components assessing discrete aspects of a syllabus. Each syllabus component could receive a score from 0 to 3. The sum of component scores produces a total score of 0 to 18. The committee expected that total scores would generally inform the institution about the prevalence of diversity outcomes in the academic curriculum while disaggregating component scores by department and subject would result in findings that can help the institution address the consistent delivery of diversity-related outcomes in different aspects of a course.

METHODS

All course syllabi (N=504) from the 2013 and 2014 academic years were reviewed by the assessment committee. Syllabi were distributed to and scored by five evaluators based on area of expertise. For example, Photography syllabi were evaluated by a faculty member from the Photography Department. Each scoring sheet required the course number, which made it possible to disaggregate findings by department (Bachelor of Art [BA], BFA, Master of Arts [MA], Master of Fine Arts [MFA]), subject (e.g., Photography, English), and academic level (Undergraduate, Graduate).

RESULTS

Normalization of Scoring

Initial analysis of syllabi grading revealed a difference in scoring among evaluators (see Table 1), a problem further complicated by faculty evaluators being limited to grading syllabi within their realm of expertise. This meant that scores for subjects ended up correlating to the evaluator. To account for the severity (tendency to score harshly) and leniency (tendency to score less harshly) biases of evaluators, institutional research transformed component scores into z-scores.

Table 1

Syllabi Scores by Evaluator

Eval	Areas	# Syllabi	Avg. of Original Syllabus Score
1	BA, MA (all subjects)	173	13.2
2	BFA (AT, FM, PA, SC)	139	7.2
3	BFA (NG, PH, PR)	88	12.3
4	BFA (CP, IN), MFA (all subjects)	92	9.5
5	MFA (all subjects)	12	6.1

Converting the original scores into z-scores resulted in a distribution with mean 0 and variance 1 for each evaluator. Then, all component z-scores were subsequently transformed back to the original scale with a mean of 1.5, and desired variance was adjusted for the highest scores to be as close as possible to an upper limit of 3. This upper bound on component scoring also allows for a highest possible total score of 18, which is the upper bound of the original total score.

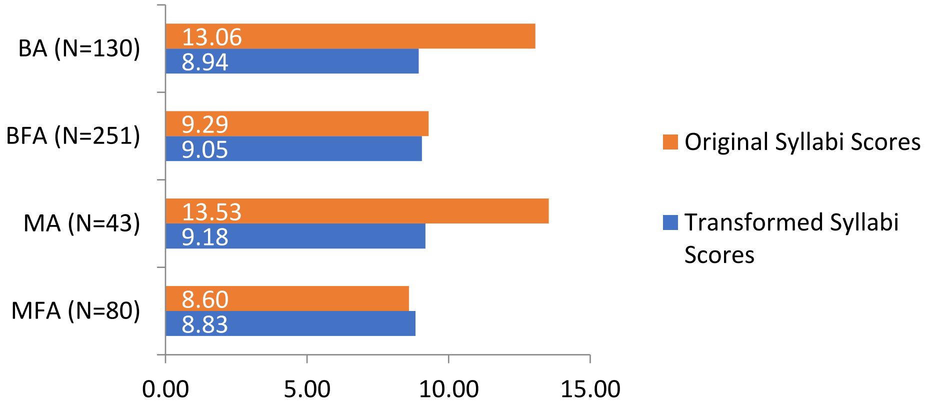
Thus, the method of twice-transforming scores allows for normalization within evaluators, which could also be interpreted using the original component and total score scales.

Assigning syllabi based on area of expertise also caused a disproportion of syllabi reviewed among evaluators—note that evaluators #1 and #2 evaluated 173 and 139 syllabi, respectively, accounting for 62% of all evaluated syllabi while the rest were distributed among the other three evaluators.

Syllabi and Component Scores Across Departments

There were four academic departments at SFAI at the time of the study: BA, BFA, MA, and MFA. These departments correspond to the degrees offered by the college. Figure 1 shows the original and transformed total scores per department.

Figure 1. Total Scores by Department. The scoring scale is 0–18.



Based on original total scores, the BA program (13.06) is satisfying the Diversity Rubric better than the BFA program (9.29). However, transformed total scores show the opposite; the BFA program scores 9.05 against the Diversity Rubric while the BA program scores 8.94. It is clear when reviewing transformed total scores by department that syllabi from the MFA department show the least evidence of satisfying diversity-related outcomes.

The transformed scores, which should have eliminated scoring tendencies by evaluators, will be used from this point forward for both Component Scores and Syllabi Scores; assume that the transformed scores are being used in subsequent charts. Keep in mind that the normalization has suppressed variance in scoring, so differences can be scrutinized up to the hundredths.

Figure 2. Component Scores by Department. The scoring scale is 0–3.

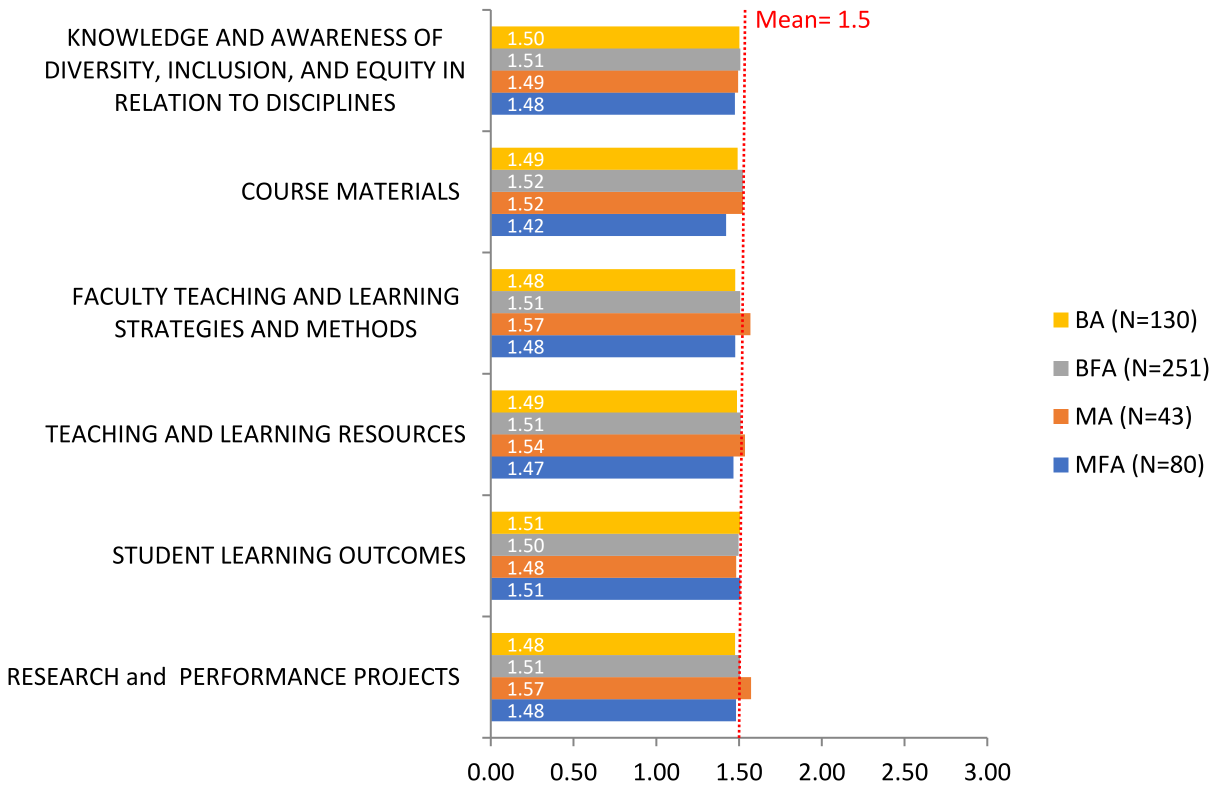
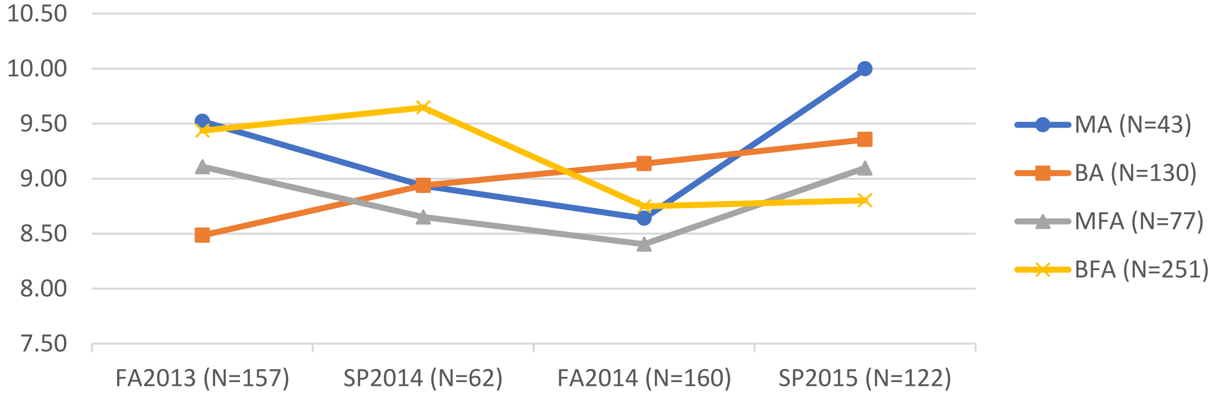


Figure 2 allows the committee to see how each department (BA, MFA, MA, MFA) fares in each syllabi subcomponent in relation to one another. Because department chairs approve courses in their purview, it may help to see how some departments better evidence diversity outcomes in particular syllabi components. For example, the MA department scores significantly better in Research and Performance Projects (1.57 compared to scores of 1.48 [BA], 1.51 [BFA], and 1.48 [MFA]) than the other departments; perhaps a review of the MA syllabi might help the other departments better address this syllabus component. At the same time, the chart shows that the MFA department is particularly poor (1.42 compared to 1.49 [BA], 1.52 [BFA], and 1.54 [MA]) in evidencing diversity outcomes in Course Materials.

Figure 3. Total Scores by Term and Department. The scoring scale is 0–18.



The committee also wanted to know how syllabi scores have fared since the institution defined diversity for the campus in 2012. Figure 3 shows among departments, MA and BA syllabi show improvement in scoring against the Diversity Rubric from Fall 2013 to Spring 2015, though only BA syllabi show a steady upward trend. Faculty representing the BA department indicated developing Mathematics and Science courses during this period to better address diversity outcomes. Perhaps the above chart shows that, with attention, a department can improve in evidencing diversity outcomes in syllabi over time.

Figure 4. Component Scores, All Departments (N=504).

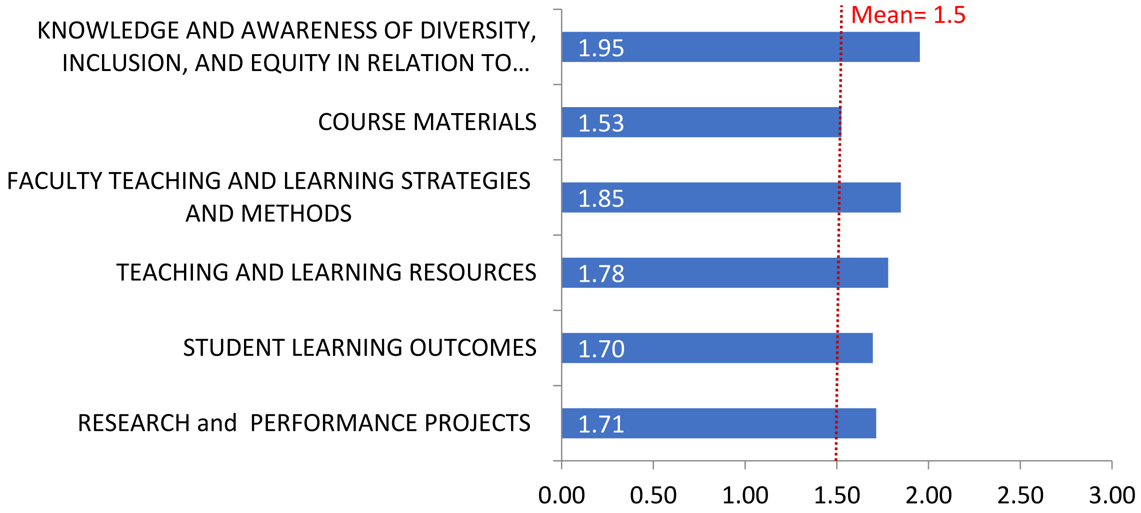
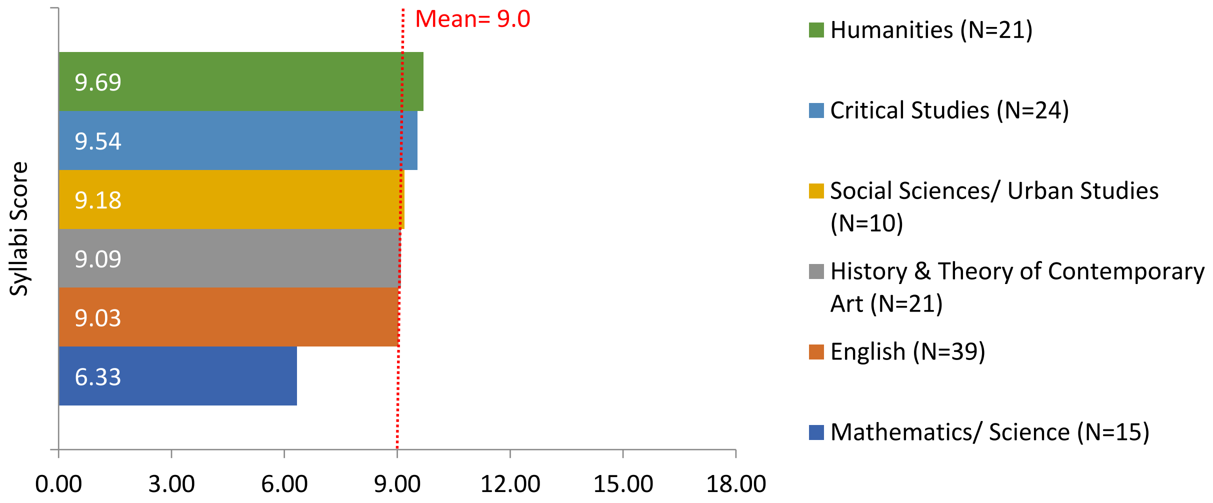


Figure 4 shows across departments, diversity outcomes in syllabi components are weakest in Course Materials (1.53) followed by Student Learning Outcomes (1.70) and Research and Performance Projects (1.71).

Total component scores will now be assessed by department at the subject level to determine if syllabi of particular subjects struggle to evidence diversity outcomes.

Figure 5. Total Scores by BA Subject. The scoring scale is 0–18.

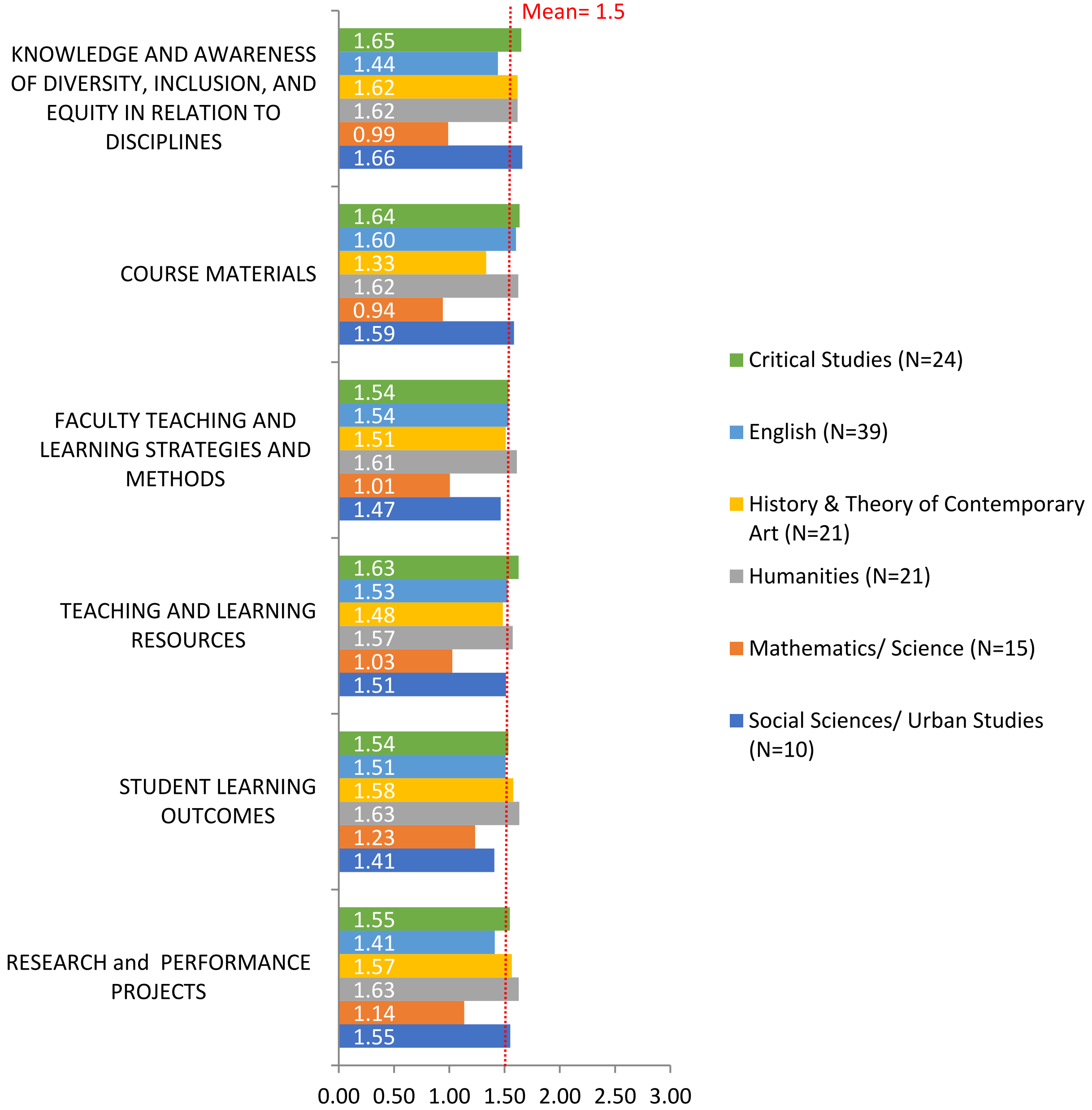


BA Department Results

Figure 5 shows within the BA department, Mathematics/Science syllabi score lowest on diversity (6.33) compared to other subjects in the department. On the other hand, Humanities (9.69) and Critical Studies (9.54) syllabi score significantly above the mean.

A review of the component scores within the BA program (see figure 6) allows identification of areas where diversity outcomes are deficient or plentiful. For Mathematics/Science, the deficiency is evident in Course Description & Orientation (0.99), Course Materials (0.94), Learning Strategies and Methods (1.01), and Teaching & Learning Resources (1.03).

Figure 6. Component Scores by BA Subject. The scoring scale is 0–3.



Figures 5 and 6 were reviewed by the faculty members of the Program Assessment Committee. As previously mentioned, a faculty representative who had written the BA and MA self-study the prior year mentioned that a concentrated effort had been made for Mathematics/Science courses to better address diversity outcomes. Figure 7 shows an increase in inclusion of diversity in the Mathematics/Science program over the course of four terms with a slight decrease in Spring 2015.

Figure 7. Total Scores by Term and BFA Subject. The scoring scale is 0–18.

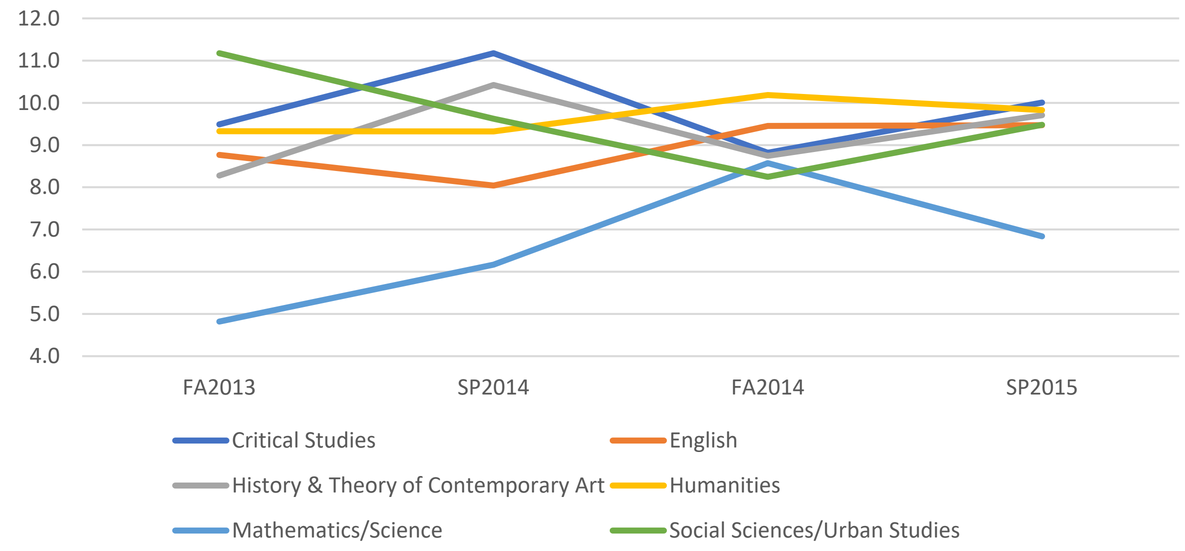
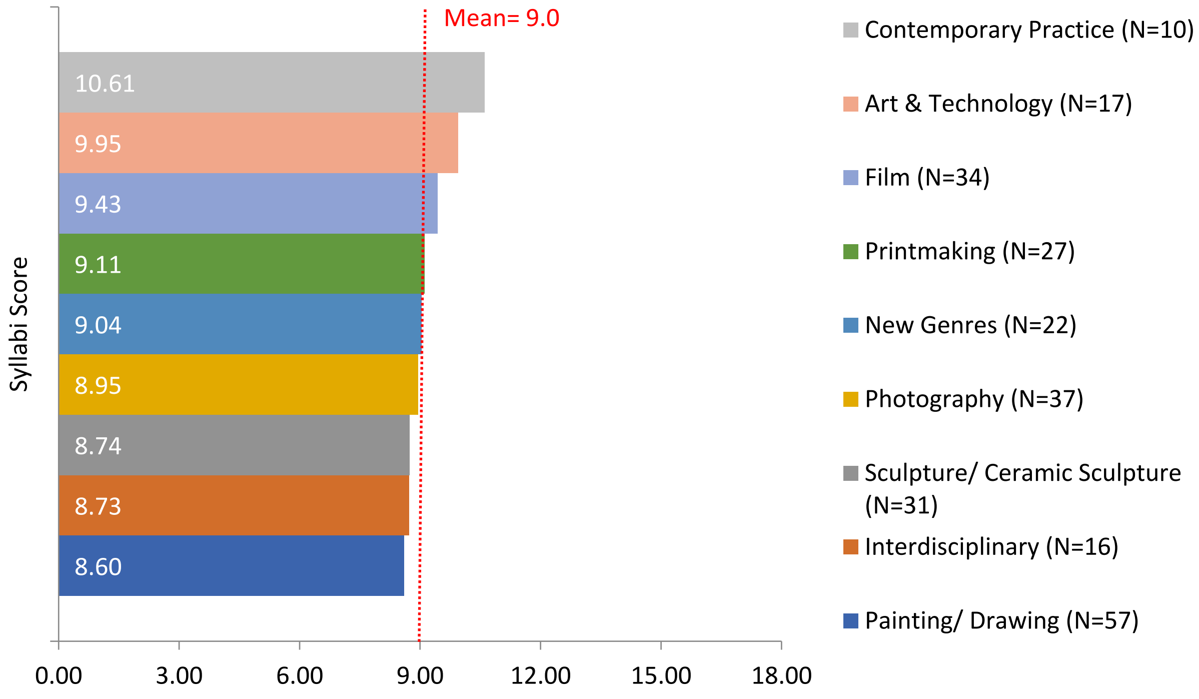


Figure 8. Total Scores by BFA Subject. The scoring scale is 0–18.



BFA Department Results

Figure 8 shows among BFA subjects, Contemporary Practice syllabi (10.61) scored significantly above the mean for diversity outcomes followed by Art & Technology (9.95) and Film (9.43). Contemporary Practice is the freshman core class for degree-seeking undergraduates at SFAI; this foundation class for incoming students appears to have been designed to be particularly sensitive towards diversity.

Figure 9. Component Scores by BFA Subject. The scoring scale is 0–3.

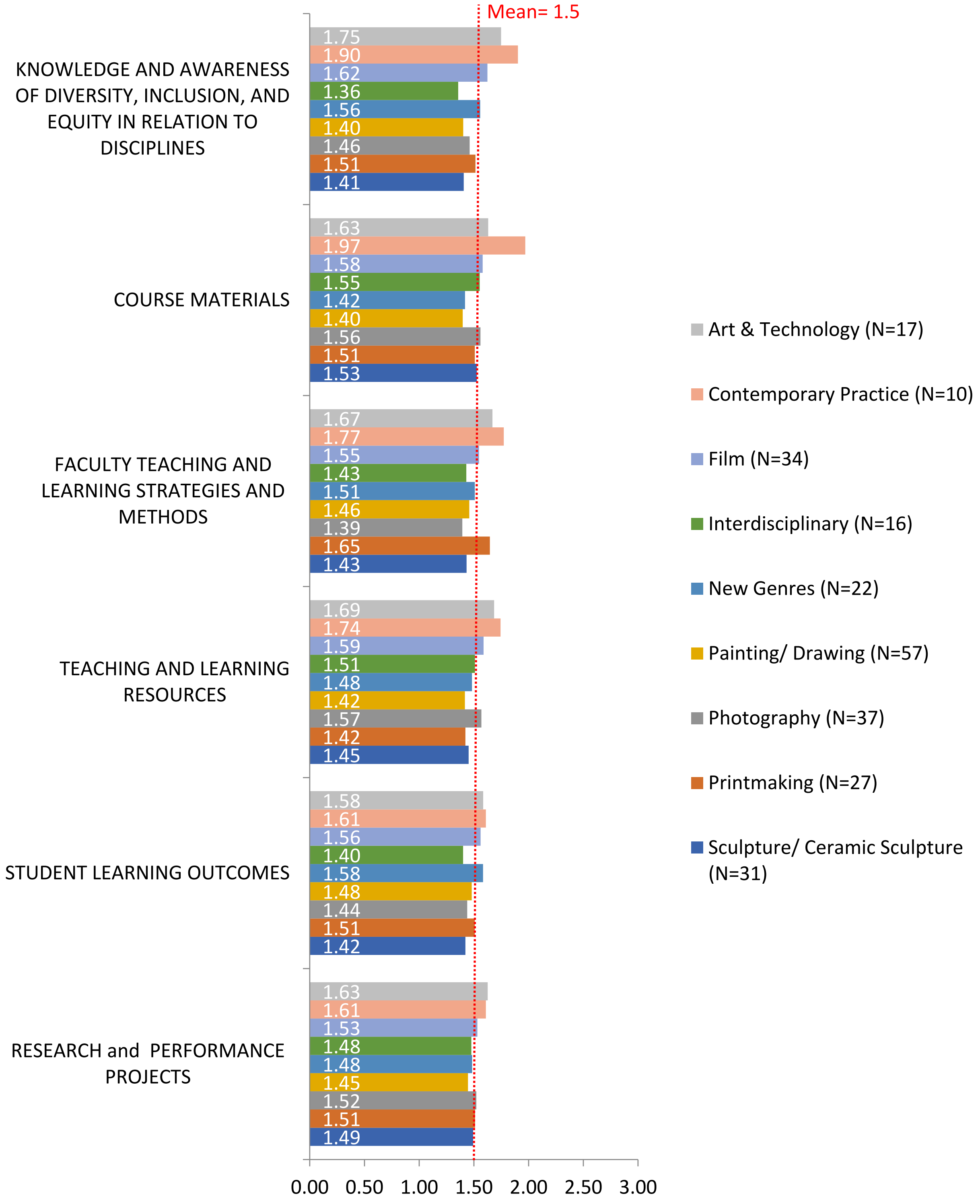


Figure 9 shows Component Scores of BFA subjects. These scores help identify which subjects are more advanced in including diversity in the curriculum. Subjects with higher component scores may be examples of how to include diversity outcomes in syllabi. For example, Art & Technology course syllabi may serve to illustrate how diversity can be articulated in Research and Performance Projects since they have the highest score of 1.63 within the component.

MA Department Results

Figure 10 shows among MA subjects, course syllabi from Exhibition & Museum Studies (EMS; 7.31) are significantly below the mean at including diversity in the curriculum. The rest of the MA subjects have syllabi scoring above the mean. In all syllabi components, course syllabi for EMS score significantly lower than other MA subjects in evidencing diversity outcomes. This indicates a required improvement in EMS for developing syllabi to evidence diversity outcomes. Figure 11 shows that MA course syllabi are particularly weak on the following components: Knowledge and Awareness of Diversity (1.12) and Teaching and Learning Resources (1.14).

Figure 10. Syllabi Scores by MA Subject. The scoring scale is 0–18.

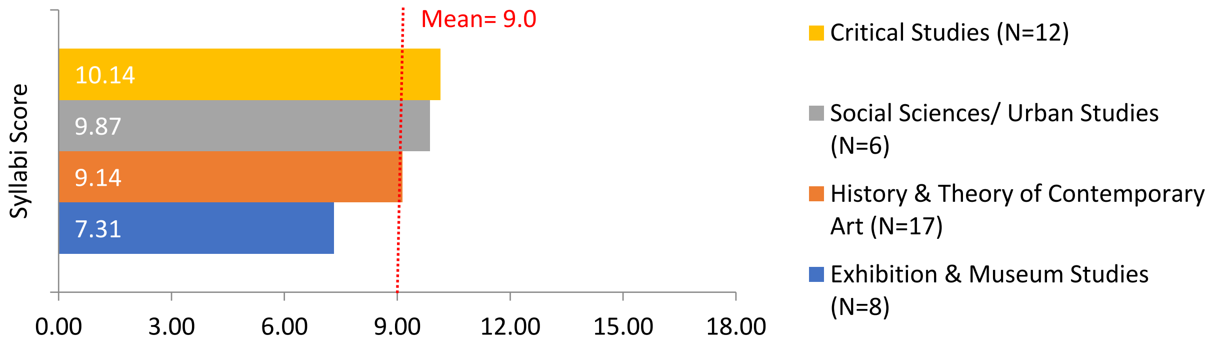
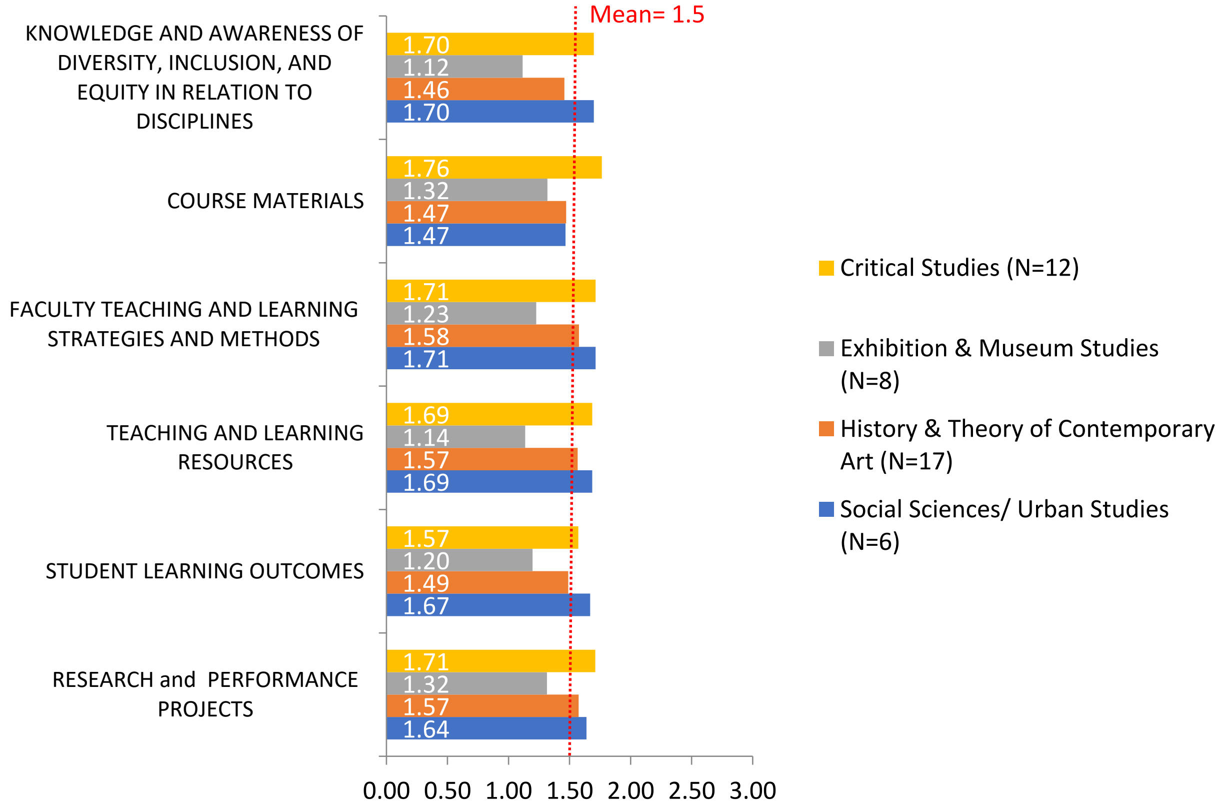


Figure 11. Component Scores by MA Subject. The scoring scale is 0–3.



MFA Department Results

The MFA program is a single subject, Studio Art (see figure 12). For the component-level evaluation, MFA subjects are consolidated into Graduate Studio (including Film, Painting/Drawing), Interdisciplinary, and Post-Baccalaureate to reflect the current curricular structure.

Figure 12. Syllabi Scores by MFA Subject. The scoring scale is 0–18.

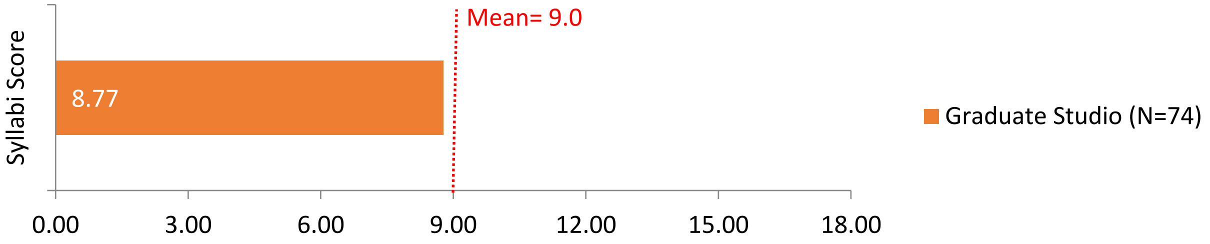
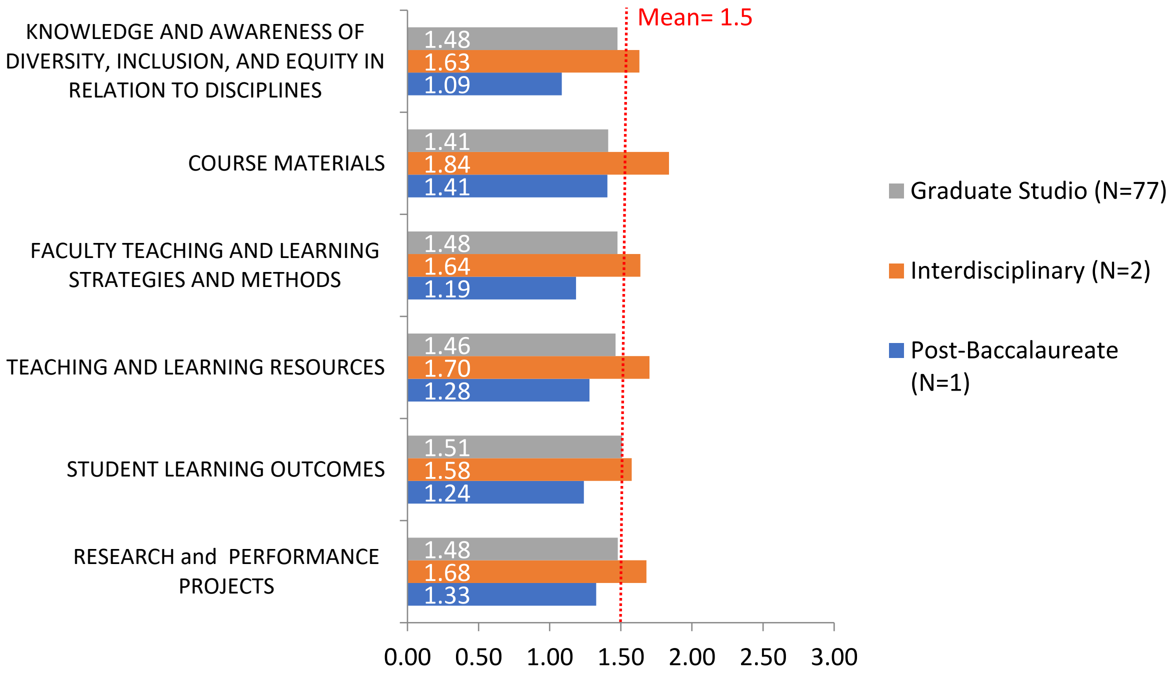


Figure 13. Component Scores by MFA Subject. The scoring scale is 0–3.



Discussion

Institutional research’s engagement in the review of syllabi against a diversity rubric resulted in the following insights for the Program Assessment Committee.

1. The process of collecting the syllabi revealed that only 57% (504/890) were available, and a handful of syllabi could not be evaluated due to being incomplete. Therefore, the committee recommended a syllabi checklist and screening process for both compliance and assessment—that is, every course should have a syllabus, and every syllabus should be reviewed to at least fit the college's syllabus template if not the college's diversity standards.
2. Evaluators, especially when assigned to score content areas in which they have expertise, might be more lenient or severe in their ratings.

Therefore, it is advisable to randomly and equally distribute syllabi among evaluators regardless of area of expertise. This will minimize the effects of lenient or rigid scoring; additionally, evaluating syllabi outside of one's area of expertise will show the differing ways in which subject areas have addressed diversity outcomes.

3. Course Materials scored the lowest (1.53) for the presence of diversity-related outcomes, which is significantly less than the other components, whose scores ranged of 1.70 to 1.95. This score may be affected by classes in which course materials are not stated in syllabi because they are determined throughout the semester based on individual student projects. For example, an instructor of a graduate tutorial class will refer to the work of a particular artist or piece of critical literature in relation to student progress. Faculty at SFAI are aware that this method of teaching occurs in a significant number of courses: Undergraduate Tutorials, Senior Seminars (a capstone class for BFA candidates), Graduate Tutorials, and Graduate Critique Seminars. For MFA candidates, these courses comprise 40% of the degree requirements.

4. Applying the diversity rubric to course syllabi was one metric to understand coverage of diversity in a classroom; however, the method does not assess the student experience of learning through a diversity-informed curriculum. A more complete assessment of diversity outcomes in the curriculum would require both a syllabi audit and an assessment of diversity-related questions from course evaluations or a diversity-specific survey for students based on classroom experience.

Conclusion

Institutional research facilitated assessment-related research helpful in understanding how diversity was being enacted in the curriculum pursuant to the institution's Diversity Statement. Institutional research also worked with faculty members who played a key role by creating the Diversity Rubric and with other Academic Affairs staff, participating as evaluators of course syllabi. The study is an example of a collaborative effort wherein faculty and Academic Affairs staff provided input based on curricular knowledge while institutional research provided analytic expertise by identifying scoring variance among evaluators and normalizing scores and provided standardized results for decision-making.

Since the study has been distributed and presented to various stakeholders on campus, the following changes have occurred: 1) course evaluations now have a shared set of questions accounting for diversity and inclusivity; 2) it has been recommended that the school repeat the study; if repeated, the recommendation for random and equal distribution of syllabi to evaluators would be addressed. The institution is still looking for ways to address syllabi compliance through a checklist and screening process.

Appendix: SFAI Diversity Rubric for Syllabi

The SFAI Diversity Rubric for Syllabi was written to be applicable to all types of courses and may be useful for other institutions. The scoring sheet requires a course number and semester, which allowed institutional research to aggregate results by department, by subject, and through time.

Course Name:		Course Number:	Semester:
Components	Stage One: Emerging	Stage Two: Developing	Stage Three: Transforming
KNOWLEDGE AND AWARENESS OF DIVERSITY, INCLUSION, AND EQUITY IN RELATION TO DISCIPLINES (course description and orientation)	Few aspects of the course recognize how multiple ways of knowing impact teaching and learning in the classroom.	Many aspects of the course recognize multiple ways of knowing and incorporate these into teaching and learning practice.	Most aspects of the course incorporate multiple ways of knowing into teaching and learning practices
COURSE MATERIALS (study materials, visual archives, etc.)	Coursework as it is currently constituted is only minimally related to diversity and inclusiveness.	The value of diversity, inclusion and equities is evidenced in the course materials in certain areas and not in others. A commitment to diversity, inclusion, and equity has an informing influence, albeit inconsistently.	Evidence of a strong value for diversity, inclusion, and equity is easily apparent throughout the course materials. A commitment to diversity, inclusion, and equity clearly has an informing influence.
FACULTY TEACHING AND LEARNING STRATEGIES AND METHODS	The instructor has integrated few teaching and learning approaches designed to respond to the diverse experiences of students in their classes.	The instructor has integrated a limited, but purposeful, variety of inclusive teaching and learning approaches designed to respond to the diverse experiences of students in their classes.	The instructor has clearly integrated a variety of inclusive teaching and learning approaches designed to respond to the diverse experiences of students in their classes.
TEACHING AND LEARNING RESOURCES (Teaching and learning centers, mentoring programs, etc.)	The syllabus reveals few if any resources to support the development of inclusive teaching and learning approaches designed to respond to the diverse experiences of all students in any given classroom.	The syllabus reveals some resources to support the development of inclusive teaching and learning approaches designed to respond to the diverse experiences of all students in any given classroom.	The syllabus clearly reveals many resources to support the development of inclusive teaching and learning approaches designed to respond to the diverse experiences of all students in any given classroom.
STUDENT LEARNING OUTCOMES	Few student learning outcomes identify the need for diversity, inclusion and equity as part of their typical assessment practices.	Some student learning outcomes focus on diversity, inclusion, and equity as part of their typical assessment practices.	Most student learning outcomes focus on diversity, inclusion, and equity as part of their typical assessment practices.

RESEARCH and PERFORMANCE PROJECTS	Few course research and performance requirements reflect a commitment to diversity, inclusion, and equity as an integral asset to disciplinary and institutional integrity in form and content.	Many course research and performance requirements reflect a commitment to diversity, inclusion, and equity as an integral asset to disciplinary and institutional integrity in form and content.	Most course research and performance requirements reflect a commitment to diversity, inclusion, and equity as an integral asset to disciplinary and institutional integrity in form and content.
TOTAL NUMBER OF DESIGNATIONS IN EACH COLUMN (Number of descriptions circled above)	Add the points in this column. Each designation in this column is worth 1. _________	Add the points in this column. Each designation in this column is worth 2. _________	Add the points in this column. Each designation in this column is worth 3. _________

STANDARDS-BASED SCORE	Total Score: _____ /18
	Overall Assessment:
	see right for scoring range
		transforming = 14–18
		developing = 11–13
		emerging = 8–10
		fails to adequately satisfy = 0–7	

Using Predictive Analytics to Identify “At-Risk” Student-Athletes

Heidi Carty, Galina Belokurova
University of California San Diego

Abstract

This study illustrates how educational institutions can incorporate predictive modeling to identify incoming student-athletes who are likely to struggle academically. The study utilizes student data from an R1 (Doctoral University – Carnegie Classification) public university in California and applies a range of machine-learning algorithms (CHAID, C5, Logistic Regression, Quest, C&R Tree, Random Tree, Decision List, Neural Network, and their automated and user-created ensembles) to create a predictive model and then evaluate the model using a testing dataset. The results indicate that measures of academic competence (entrance exam performance and cumulative high school GPA) and socioeconomic factors (feeder high school rank, first-generation status, and low income) are significant determinants while student-athlete status is not a significant predictor of earning a lower GPA during one’s first term in college. The scoring models obtained identify at-risk students among both athletes and non-athletes and, therefore, are applicable more broadly. Institutional researchers collaborated with the university’s athletic department to help deploy this scoring tool for use in determining at-risk student-athletes for referral to an academic support program.

INTRODUCTION

Can one accurately identify academically at-risk athletes through predictive analytics? It is a common assumption that collegiate sports involvement may increase the likelihood of lower academic achievement because student-athletes spend considerable time training and miss classes more often than non-athletes (Watt & Moore, 2001). However, the relationship between sports and educational attainment or student success is likely more complex and varied (Ferris, Finster, & McDonald, 2004; Gayles, 2009; Richard & Aries, 1999; Comeaux & Harrison, 2011; Gayles & Hu, January 2009; Harshaw & NC DOCKS, 2009).

The definition of student success, type of educational institution, constituency, and competitiveness of athletics programs may play a critical role in whether one sees any difference in the academic achievement of student-athletes vs. non-athletes. For instance, universities with more selective admission policies graduate both athletes and non-athletes at higher rates than those with open admission. At the same time, some research has shown that involvement in athletics may still negatively affect graduation rates of athletes compared to their peers in the same institution (Ferris et al., 2004). Other studies argue that the correlation between participation in athletics and lower grades may be spurious and related to less rigorous academic preparation of student-athletes, not their participation in sports (Richard & Aries, 1999).

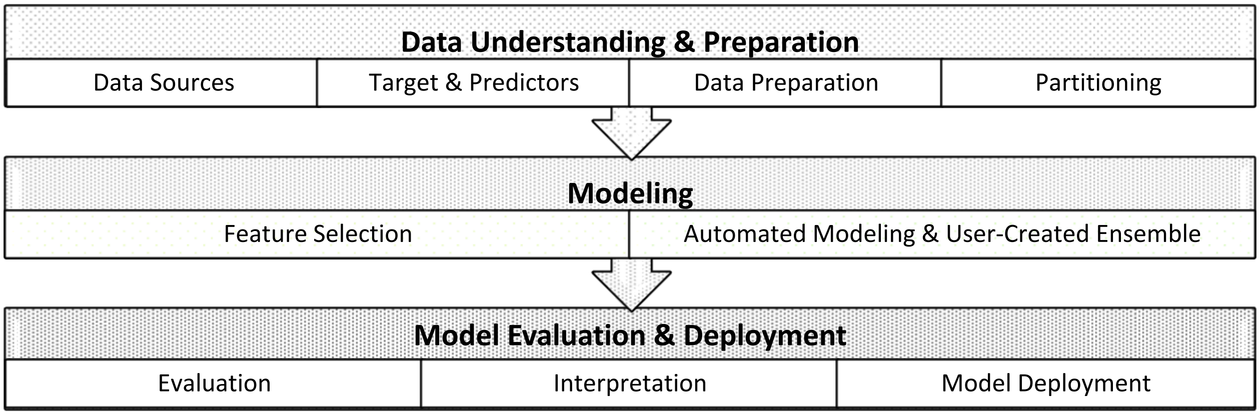
Currently, most researchers acknowledge that sports’ impact on student-athletes can be differential, have many inputs outside the university environment, and develop over time (Beron & Piquero, 2016; Comeaux & Harrison, 2011; Gayles & Hu, January 2009; Harshaw & NC DOCKS, 2009). For example, less competitive Division III athletics programs may have a more positive overall influence compared to higher-stakes Divisions I and II. Robst and Keil (2000) find that Division III athletics involvement does not limit students' academic performance. They explicitly account for student-athletes’ academic ability (e.g., SAT scores and feeder high school rank), eliminating the difference in average GPA between athletes and non-athletes. Beron and Piquero (2016) analyzed a survey of NCAA student-athletes and did not find that Division I student-athletes underperform academically compared to their Division III peers. The balancing of academics and athletics can be more difficult for different groups in terms of race and gender (McDougle & Capers, 2012; Sellers & Kuperminc, 1997).

Nevertheless, this project shows that institutions can incorporate predictive modeling into their enrollment procedures and assist their athletic departments in developing optimal programs to increase student-athletes’ academic potential.

METHODS

Any predictive modeling process consists of several steps, including defining the research question and incorporating the existing research data understanding and preparation, modeling, and model evaluation and deployment (Figure 1). This section focuses on data and modeling; the Results and Discussion sections cover model assessment and implementation.

Figure 1. Building and Deploying Predictive Models.



Data Understanding & Preparation

The data come from an R1 public university in California and include a population of the first-time first-year students who entered between 2013 and 2016 (N=26,887). The institution's student body splits approximately equally between men and women; racial/ethnic composition includes about 28 percent underrepresented minorities, 46 percent Asian, 15 percent white, and 11 percent with no racial/ethnic information available. The school has a growing Division II athletics program, and about 3 percent (N=784) of the first-time first-year students participate in collegiate athletics. Table 1 shows data used to predict being academically at risk. At risk is defined here as student-athletes achieving an end-of-first-term GPA lower than 2.6.

The predictive analytics (or machine-learning) approach partitions the original data sample into training and testing subsamples, fits the model using the training subsample, and finally applies the resulting routine to see if the groupings obtained through the model application correspond to the actual distribution found in the testing subsample. One must partition the testing from the training subsample before any analysis takes place as the testing data provide the basis for model evaluation. Training the model on the testing subsample would render the testing data useless.

Table 1
Predictor Variables

Predictor Groups		Predictor Variables
Demographic and socioeconomic		Sex at birth, family income group, student's home location, first-generation status
Institutional		Major at admission, high school ranking, college within the university applied to
Course performance in high school		Honors courses, American history requirement, total number of math courses, total number of science labs, total number of elective courses, total number of AP courses taken
Aggregate academic performance		Cumulative high school GPA, academic index score (derived by the university's admission department from a series of test scores and high school GPA)
Entrance exams		Official SAT math, verbal, & writing scores; ACT scores; number of AP tests taken with passing score
Athletic status		Involvement in intercollegiate sports

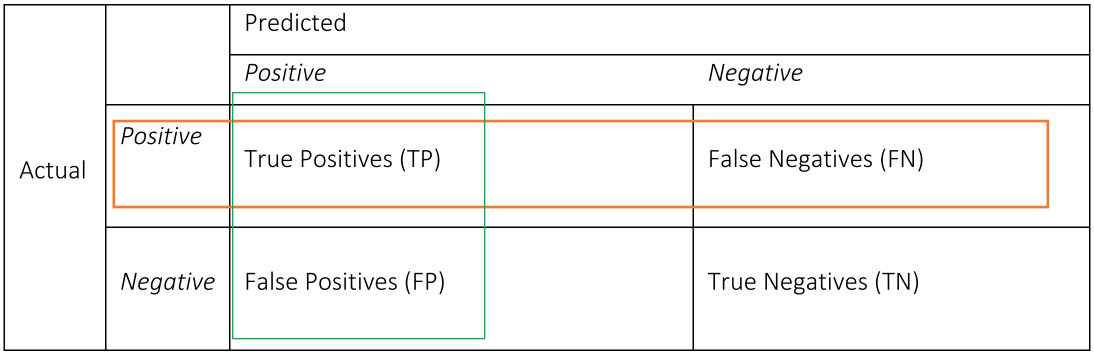
The process of fitting the model involves iterative splits of the training sample into preset categories based on predictors to approximate the distribution between classes found in the training sample. During data preparation, a screening process dropped fields missing more than 50 percent of values as well as outliers based on robust mean and standard deviation calculated, assuming that the share of outliers does not exceed 5 percent. Continuous variables went through the z-score transformation. Then, the highly associated categorical variables were merged into one based on how closely their categories predicted the target (IBM SPSS Modeler 17 Algorithms Guide, p. 13). Fifty percent of randomly selected cases then formed the training subsample, and the remaining 50 percent made up the testing subsample. The training subsample additionally underwent balancing, an artificial simulation of additional cases with the end-of-term GPA lower than 2.6 to offset the misbalance between the number of students in good standing and those with a low GPA. Balancing is especially crucial for increasing model recall.

Modeling

The model predicts a binary outcome (students at-risk/not at-risk using classifiers – supervised learning techniques relying on predefined categories). CHAID, Quest, C&R Tree, Random Tree, C5, Logistic Regression, and Decision List are the most common algorithms used with binary targets. CHAID, the best algorithm in this study, initially emerged as a market segmentation tool (Kass, 1980; IBM SPSS Modeler 17 Algorithms Guide, p.73). For categorical variables, CHAID uses a chi-squared test to split the original training sample recursively into progressively more homogenous groups using predictor variables. When isolated predictors cannot separate one target outcome from another (put at-risk and not at-risk students in separate bins), this algorithm starts merging predictors until homogeneity of the resulting groups cannot increase any further.

Upon completing the training stage, one obtains a set of rules called an algorithm that can sort an unseen-before subset of data into categories (e.g., at-risk/not at-risk) based on combinations of predictors. A confusion matrix (Figure 2) compares the model estimates with actual estimates in the testing subsample. There are three metrics used to assess the accuracy of machine-learning models, precision, recall, and overall accuracy.

Figure 2. True Positives/Negatives and False Positives/Negatives.



		Predicted	
		Positive	*Negative*
Actual	*Positive*	True Positives (TP)	False Negatives (FN)
	Negative	False Positives (FP)	True Negatives (TN)

(1) Precision refers to the share of true positive outcomes (e.g., at-risk students who were actually at risk) in the total number of results identified as positive by the model (e.g., at-risk students as determined by the model): True Positive × 100 / (True Positive + False Positive). In terms of the current model, precision is the percentage of all students classified as at risk by the model in the at-risk category after the first term. Figure 2 shows the values used in this calculation in a green box.

(2) Recall is the percentage of the actual positive outcomes (e.g., at-risk students) identified by the model in the total number of true positives (e.g., students at risk after the first term): True Positive × 100 / (True Positive + False Negative) or the percentage of the students predicted as at risk in all students actually at risk. Table 2 shows the values used in this calculation in an orange box.

(3) (True Positives + True Negatives) × 100 / (True Positives + False negatives + False Positives + True Negatives). It is the percentage of at-risk students correctly identified and not at-risk students correctly identified by the model over the total number of all students.

Table 2

Feature Selection Results

Rank	Description	Type	Importance	Value
1	Official combined SAT score	Nominal	0	1
2	High school quintile rank	Nominal	0	1
3	Official SAT Writing score	Continuous	0	1
4	Official SAT Math score	Continuous	0	1
5	Official SAT Verbal score	Continuous	0	1
6	High school GPA weighted, unlimited honors	Continuous	0	1
7	First-generation (both parents have no bachelor's degree)	Flag	0	1
8	Income group	Nominal	0	1
9	Number of AP courses taken	Continuous	0	1
10	First-generation & low-income students	Flag	0	1
11	Number of honors seminar courses in 10th & 11th grades	Nominal	0	1
12	Academic index score	Continuous	0	1
13	Home location (fill in the groups)	Nominal	0	1
14	Total number LabSci courses 9-12	Continuous	0	1
15	Total number A-G courses official	Continuous	0	1
16	Total number Math courses 7-12	Continuous	0	1
17	Department chosen at admission	Nominal	0	1
18	College with the university at admission	Nominal	0	1
19	American history requirement met?	Nominal	0	0.9999
20	Total AP courses taken/planned	Continuous	0	0.9848
21	Sex at birth	Nominal	0	0.9737
22	Number of honors sem. courses 10th grade	Continuous	1	0.945
23	Number of honors sem. courses 11th grade	Continuous	2	0.2979
24	Total number elective courses 9-12	Continuous	2	0.0285

Note: Importance of 0 refers to the most important predictors

(4) (True Positives + True Negatives) × 100 / (True Positives + False negatives + False Positives + True Negatives). It is the percentage of at-risk students correctly identified and not at-risk students correctly identified by the model over the total number of all students.

(5) The target variable identifies students not at risk (e.g., GPA 2.6 or higher) versus those at risk (e.g., less than a 2.6 GPA). The GPA threshold of 2.6 derives from the request of the athletic department in selecting student-athletes for the academic support program.

(6) All models applied included the same set of predictors characterizing students' demographic and socioeconomic backgrounds, the positions of their high schools within the state ranking, departments and majors at admission, prior academic performance, entrance exams, and student-athlete status. Table 1 displays a list of predictors.

(7) Feature selection algorithms assess variables' predictive power by running a classifier built with each predictor separately (IBM SPSS Modeler 17 Algorithms Guide p.153). The predictor rankings in a mixed model sort categorical fields according to the p-values based on Pearson's chi-squared and continuous variables according to p-values based on the F statistic. Table 2, Feature Selection Results, shows the ranking in which the p-values are transformed on the scale from zero (least) to one (most significant). Importance of zero corresponds to the most significant predictors.

(8) The feature selection helps researchers reduce the number of predictors early in the process to make the interpretation clearer later. The measures of academic competence (entrance exam performance and cumulative high school GPA) as well as socioeconomic background (feeder high school rank and first-generation and low-income group) comprise the top 10 most important predictors. At this stage, being a student-athlete does not appear to be a significant predictor, indicating that participation in collegiate sports does not impair students' first-term GPA.

Table 3

Top Five Most Important Predictors by Algorithm

Top 5 Predictors / Model	Automated	C5	CHAID	Logistic	Neural Net
Predictor 1	Academic index score	Academic index score	High school quintile rank	Official combined SAT score	Total AP courses taken/planned
Predictor 2	Official combined SAT score	High school quintile rank	Official combined SAT score	High school quintile rank	High school GPA weighted, unlimited honors
Predictor 3	SAT Writing score	SAT Writing score	SAT Math score	SAT Writing score	SAT Writing score
Predictor 4	High school quintile rank	Department at admission	SAT Writing score	SAT Math score	Number of AP courses taken
Predictor 5	SAT Math score	First generation	SAT Verbal score	High school GPA weighted, unlimited honors	Academic index score

RESULTS

The automated model-selection procedure in SPSS Modeler (IBM SPSS Modeler 17) rated CHAID, Logistic Regression, and Decision List as the top three with overall accuracies of 84.1, 84.0, and 44.8 percent, respectively. The Decision list does not perform as well as CHAID and Logistic Regression. A custom ensemble was created using four algorithms—CHAID, C5, Logistic Regression, and Neural Network—to improve model recall.

Table 3 shows general agreement among four of the algorithms on the top five predictors produced through different procedures. The most robust predictors of being academically at risk are academic index score, high school quintile rank, and SAT scores. However, the Neural Net algorithm, an algorithm intended to uncover patterns hidden in the data without any input from the researcher, produces a different list of top five predictors emphasizing advanced placement courses.

Most of the students do well academically, and the model is likely to identify them with high accuracy.

Table 4 shows that the automated model identified close to 11 percent of students at risk in the testing sample (model recall—in green), and, within that, 60 percent correctly appeared in the at-risk category (model precision—in red). The custom ensemble increases the model precision to 69 percent but at the cost of dropping the model recall to a mere 4 percent. At the same time, model specificity, in blue, is almost 99 percent (i.e., students in good academic standing correctly appeared in the not-at-risk group) for both models.

Table 4

Automatically Created Model vs. Custom Ensemble

		Automated Model Selection		Custom Ensemble	
Fall GPA Category (rows)	Predicted Fall GPA Category (columns)	1=Predicted Low GPA	2=Predicted Satisfactory GPA	1=Predicted Low GPA	2=Predicted Satisfactory GPA
1=Low GPA	Count	376	3111	53	1276
	Row %	10.8	89.2	4.0	96.0
	Column %	60.5	15.2	68.8	13.2
	Total %	1.8	14.8	0.5	13.1
2=Satisfactory GPA	Count	246	17303	246	17303
	Row %	1.4	98.6	1.4	98.6
	Column %	39.6	84.8	39.6	84.8
	Total %	1.2	82.3	1.2	82.3

Note: Precision of each model is in green; Recall – red, Specificity - blue

DISCUSSION

What does this analysis say about the influence of collegiate sports on the academic performance of student-athletes? None of the algorithms, whether fitting individually or as an ensemble, featured the participation in the university athletic program as a factor. This result agrees with many studies that emphasize that the association between collegiate sports and lower GPA may be more complex (Beron & Piquero, 2016; Gayles, 2009; Richard & Aries, 1999). Furthermore, the study showed that student-athletes do not have needs fundamentally different from the rest of the student body, and academic preparation seemed to influence student success the most.

The comparison between automated and custom models also shows that modeling process depends on balancing conflicting goals of identifying as many at-risk students as possible without over-identifying and unfairly referring students to academic support programs. A higher-precision model may assist in picking a smaller group of students who are very likely to struggle academically. A higher-recall model produces a bigger group of potential referrals. Ultimately, the choice between these two approaches depends on the particular situation in each athletic department.

Despite the low recall of the resulting model, the athletics department implemented it to identify potentially at-risk students. The model was utilized in 2017 and 2018 to separate a small group of incoming first-time first-year athletes who were predicted to earn a GPA of less than 2.6. An athletic department-based program providing additional tutoring, academic advising, and time management training invited these student-athletes to participate. All of the students selected for the program participated with each earning a GPA higher than 2.6 in their first quarter. The athletic department considers the program a success and plans to continue and expand the program as needed. The collaboration between institutional research and athletics departments aimed at incorporating predictive modeling into enrollment procedures to identify at-risk student-athletes not only benefits student-athletes but also provides a successful approach to develop optimal academic support programs across campus.

REFERENCES

Beron, K. J., & Piquero, A. R. (2016). Studying the determinants of student-athlete grade point average: The roles of identity, context, and academic interests. *Social Science Quarterly, 97*(2), 142–160. https://EconPapers.repec.org/RePEc:bla:socsci:v:97:y:2016:i:2:p:142-160

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* New York: Chapman & Hall/CRC.

Comeaux, E., & Harrison, C. K. (2011). A conceptual model of academic success for student-athletes. *Educational Researcher, 40*(5), 235–245. https://doi.org/10.3102/0013189X11415260

Ferris, E., Finster, M., & McDonald, D. (2004). Academic fit of student-athletes: An analysis of NCAA Division I-A graduation rates. *Research in Higher Education, 45*(6), 555–575. Retrieved from http://www.jstor.org/stable/40197361

Fürnkranz, J. (2011). Decision list. In C. Sammut & G. I. Webb (eds.) *Encyclopedia of machine learning*. Boston, MA: Springer.

Gayles, J. G., & Hu, S. (January 01, 2009). The influence of student engagement and sport participation on college outcomes among Division I student-athletes. *The Journal of Higher Education, 80*(3), 315–333.

Gayles, Joy. (2009). The student-athlete experience. *New Directions for Institutional Research, 2009,* 33–41. 10.1002/ir.311.

Harshaw, C. E., & NC DOCKS at The University of North Carolina at Greensboro. (2009). The role of intercollegiate athletics in the retention of first-time, first-year students at NCAA Division III institutions with football IBM SPSS Modeler 17 (2015), IBM Corporation.

Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 29*(2), 119–127. DOI:10.2307/2986296

Loh, W.-Y., & Shih, Y.-S. (1997), Split selection methods for classification trees. *Statistica Sinica, 7*, 815–840.

McDougle, L. & Capers IV, Q. (2012). Establishing priorities for student-athletes: Balancing academics and sports. *Spectrum: A Journal on Black Men, 1,* 71–77. 10.2979/spectrum.1.1.71

Richard, S., & Aries, E. (1999). The Division III student-athlete: Academic performance, campus involvement, and growth. *Journal of College Student Development, 40.*

Rivest, R. L. (1987). Learning decision lists. *Machine Learning, 2*, 229–246.

Robst, J., & Keil, J. (2000). The relationship between athletic participation and academic performance: Evidence from NCAA Division III. *Applied Economics, 32,* 547–558. 10.1080/000368400322453

Sellers, R., & Kuperminc, G. P. (1997). Goal discrepancy in african american male student-athletes' unrealistic expectations for careers in professional sports.' *Journal of Black Psychology, 23*(1), 6–23.

Watt, S. K., & Moore, J. L. (2001). Who are student-athletes" In S. Watt & J. Moore (eds.) *Student Services for Athletes: New Directions in Student Services* (no. 93). San Francisco: Jossey-Bass.

"I had my first paper published in 1978. It was a study of a phone survey about why students drop their classes at a community college. In retrospect, being published was a valuable stepping stone that gave me more understanding about my role as an institutional researcher and helped me gain confidence in my ability to conduct useful analysis and research. It encouraged me to continue in IR, resulting in 40 rewarding years in institutional research. I believe *The CAIR Report* can also be that valuable stepping stone for all California IR professionals to help build a successful career in institutional research."

Robert Daly
First CAIR President

"*The CAIR Report* helps to elevate the superb work of talented institutional research professionals that is presented at the CAIR conference. This annual resource is sure to add great professional development value to new, mid-ranged and seasoned institutional researchers."

Kristina Powers
CAIR President, 2016

"The CAIR Report follows in the best traditions of CAIR by facilitating the collegial exchange of ideas of common interest to those in, and around, the field of institutional research."

Juan Ramirez
CAIR President, 2017

ABOUT THE CAIR REPORT

The CAIR Report is a publication of the California Association for Institutional Research (CAIR) featuring research shared by presenters at the annual CAIR conference in written form. Articles are selected both for their caliber of research and reflection of the conference theme. The publication also functions as a mechanism for sharing emerging research in institutional research with those who are not able to attend the conference in person.

ABOUT CAIR

The California Association for Institutional Research (CAIR) is dedicated to the fostering of unity and cooperation among persons having interests and activities related to institutional research and/or planning in California institutions of postsecondary education. CAIR provides a forum for the dissemination of information and interchange of ideas on problems of common interest in the field of institutional research.

In addition, CAIR promotes the continued professional development of individuals engaging in institutional research and fosters the unity and cooperation among persons having interests and activities related to research.

2019

California Association
CAIR
for Institutional Research