**Prospective Teacher Course Evaluation Survey Pilot Spring 2019 Analysis**

**Prepared by Rebecca Pérez and Lisa Elfring**

**UA Office of Instruction and Assessment**

**September 5, 2019**

**Background**

Teaching is a core function of the University of Arizona, but there is considerable diversity across campus in the types of evidence used to evaluate teaching effectiveness. Many UA departments draw largely or exclusively on student evaluations of teaching, locally termed TCEs. Reliance on these instruments as measures of teaching quality is problematic on several fronts, one of which is the impact of bias on students’ responses to course-evaluation items. A wealth of recent research literature concerning student evaluations of teaching (SET) demonstrates that students’ implicit biases impact their responses about the effectiveness of their instructors. **Collectively, the reports demonstrate that student bias with regard to instructors’ gender, racial and ethnic background, fluency in English, LGBT orientation, age, and physical attributes, in addition to course attributes such as meeting time and format and whether the course is elective or required, all impact students’ perceptions of the effectiveness of an instructor (Stark and Freishtat, 2014; Spooren et al, 2013; Hamermesh and Parker, 2003; MacNell et al, 2015; Clayson, 2013)**. These data have led some US universities, including the [University of Southern California](http://cet.usc.edu/resources/instructor-course-evaluation/) and the [University of Oregon](https://provost.uoregon.edu/revising-uos-teaching-evaluations) in the past year, to prohibit or limit the use of student-evaluation data in decisions about promotion, tenure, or contract renewal for faculty members and lecturers.

To address this issue, the Office of Instruction and Assessment worked with a faculty committee to develop a set of new “core” student-evaluation questions that: 1. Could be predicted to reduce the impact of implicit bias by focusing on questions related to the students’ experiences in courses; 2. Provide feedback that helps instructors to focus their efforts to improve their teaching; and 3. Can be applied to teaching in the many formats in which it takes place across the University. After iterative review, focus groups, and initial student testing, the end result is a set of 12 forced-response student-evaluation questions, shown in Table 2. The new questions were piloted in a survey in combination with the existing TCE questions in several classes in Spring 2019, and the results are described below.

**Methods**

**Participant Group**

37 instructors teaching a total of 59 courses volunteered to pilot the prospective Teacher Course Evaluation (TCE) survey. Of these, 39 courses received at least 3 student responses to the pilot TCE survey (28 instructors). For current TCE items, one course did not receive a report due to low response, and the group sample sizes vary for each question depending on whether that item was included on the TCE template for that course component. The courses included 1 studio, 1 independent study, and 37 lectures, which included 14 general-education classes. There were 11 tenured, 1 tenure-eligible, 21 career-track/adjunct and 6 teaching assistant instructors.

**Procedure**

Pilot TCE surveys were administered to classes of the volunteer instructors during the final week of the semester. Each instructor was sent a course-specific link to the Qualtrics survey. Individual instructors varied in their processes to provide incentive for student participation. Each instructor was sent a summary of their students’ ratings on the pilot TCE surveys after the semester was complete and grades were finalized. Instructors were also asked for feedback on the results of the pilot TCE and current TCE reports.

The data were analyzed using SPSS v.25. Independent t-tests were used to compare means between groups. Power analyses were completed using G-Power software. T-tests were used to compare binary groups.

**Results**

**Course and Instructor Breakdown**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1**  *Breakdown of courses in Pilot TCE* | | | | | | | | | | |
|  |  | | | Gender\* | | Mode of Delivery | | | Course Level | |
| College | Subject/  Program | Instructors | Male | Female | In-person | Fully Online | Hybrid | Undergrad | Graduate |
| CALS | 4 | 4 | 1 | 3 | 4 | 1 | 1 | 6 | 0 |
| CAPLA | 2 | 2 | 0 | 2 | 0 | 2 | 0 | 2 | 0 |
| Eller | 1 | 1 | 0 | 1 | 2 | 0 | 0 | 2 | 0 |
| FA | 3 | 3 | 0 | 3 | 4 | 0 | 0 | 4 | 0 |
| HUM | 7 | 8 | 4 | 4 | 8 | 4 | 0 | 12 | 0 |
| MEZCOPH | 2 | 2 | 1 | 1 | 3 | 0 | 0 | 2 | 1 |
| NURS | 1 | 2 | 0 | 2 | 0 | 2 | 0 | 0 | 2 |
| SBS | 5 | 5\*\* | 2 | 3 | 3 | 2 | 1 | 6 | 0 |
| Science | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |
| UA South | 1 | 1\*\* | 0 | 1 | 0 | 1\*\*\* | 0 | 1 | 0 |
| **Total:** | **10** | **27** | **28** | **8** | **21** | **25** | **12** | **2** | **36** | **3** |

\*Gender reflects the range of genders identified by instructors and reported via Analytics.

\*\*One instructor taught a course in each college; counted only once.

\*\*\*This course offered as Interactive/Webcast.

**Pilot TCE Items t-test: Gender**

Only one pilot survey item showed a significant difference between genders. On average, student responses to the pilot TCE item “The course material and activities (D2L site, assigned readings, presentations, etc.) helped me learn in this course.” were statistically significantly lower for females (*M* = 4.19, *SE* = 0.54) than for males (*M* = 4.47, *SE* = 0.25), t (35.59) = -2.21, p = 0.03 with a medium effect size of *d* = 0.73. The average for females was lower by 0.28 (95% CI, -0.02 to 0.54) than the average for males. The power of this analysis is 0.50, which increases the likelihood of a Type II error.

The other 11 items did not show a statistically significant difference between genders, as detailed in Table 2.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 2**  *Results of t-test and Descriptive Statistics for Statistically Non-significant Response Mean Differences to Pilot TCE items by Gender* | | | | | | |
| Pilot TCE Item | Group | | | | 95% CI for Mean Difference | *t*  (*df*=35) |
| Female  *(n* = 28) | | Male  *(n* = 11) | |
| Mean | SD | Mean | SD |
| The learning objectives for this course were clear to me. | 4.35 | 0.50 | 4.50 | 0.25 | -0.48, 0.16 | -0.99 |
| In this course, I was encouraged to participate through class activities, projects, and/or assignments. | 4.62 | 0.28 | 4.48 | 0.37 | -0.08, 0.36 | 1.30 |
| This course expanded my knowledge and skills in this subject matter. | 4.37 | 0.51 | 4.53 | 0.29 | -0.50, 0.17 | -1.00 |
| I was treated with respect in this course. | 4.58 | 0.31 | 4.75 | 0.19 | -0.37, 0.04 | -1.65 |
| I was encouraged to analyze and/or apply the concepts and skills taught in this course. | 4.48 | 0.37 | 4.57 | 0.17 | -0.26, 0.09 | -1.00 |
| This course helped me to connect the skills I learned to the world around me. | 4.31 | 0.48 | 4.41 | 0.30 | -0.35, 0.17 | -0.72 |
| I feel I learned the subject matter well enough to help another student in this course. | 4.12 | 0.57 | 4.24 | 0.39 | -0.50, 0.25 | -0.66 |
| Course presentations, materials, procedures, and deadlines were clearly organized. | 4.18 | 0.68 | 4.47 | 0.25 | -0.73, 0.13 | -1.41 |
| I regularly/frequently had the opportunity to ask questions about concepts and skills in this course. | 4.42 | 0.33 | 4.42 | 0.30 | -0.23, 0.23 | -0.04 |
| I received feedback on my course work/assignments throughout the semester. | 4.43 | 0.31 | 4.34 | 0.34 | -0.13, 0.33 | 0.87 |
| I received feedback on course work/assignments that helped me learn. | 4.26 | 0.45 | 4.33 | 0.37 | -0.38, 0.25 | -0.42 |

**Current TCE Item t-test: Gender**

There were five current TCE items with statistically significantly higher response means for males than females, and one item with a statistically significantly higher response mean for females.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 3**  *Results of t-test and Descriptive Statistics for Statistically Significant Response Means Differences in TCE items by Gender* | | | | | | | | | | |
| TCE Item | Group | | | | | | 95% CI for Mean Difference | *t* | *df* | *p* |
| Female | | | Male | | |
| Mean | SD | n | Mean | SD | n |
| TCEI2. Usefulness of in-class activities | 4.03 | 0.62 | 25 | 4.43 | 0.29 | 11 | -0.80, -0.04 | -2.05 | 34 | .05 |
| TCEI4. Usefulness of texts and readings | 3.97 | 0.65 | 25 | 4.39 | 0.22 | 11 | -0.72, -0.13 | -2.90 | 32.85 | .01 |
| TCEI7. Amount Learned | 3.74 | 0.55 | 27 | 4.00 | 0.21 | 11 | -0.51, -0.01 | -2.10 | 35.96 | .05 |
| TCEI11. Critical Thinking | 4.15 | 0.51 | 25 | 4.43 | 0.18 | 11 | -.052, -0.05 | -2.48 | 33.16 | .02 |
| TCEI18. GPA | 4.28 | 0.44 | 27 | 4.01 | 0.14 | 11 | 0.07, 0.45 | 2.80 | 34.97 | .01 |
| TCEI23. Instructor comparison | 3.74 | 0.78 | 18 | 4.24 | 0.33 | 8 | -0.95, -0.05 | -2.28 | 24.00 | .01 |

The remaining 18 current TCE items did not show statistically significant differences in response means between males and females.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 4**  *Results of t-test and Descriptive Statistics for Statistically Non-significant Response Means Differences in TCE items by Gender* | | | | | | | | | | | |
| TCE Item | Group | | | | | | 95% CI for Mean Difference | *t* | *df* | *p* |
| Female | | | Male | | |
| Mean | SD | n | Mean | SD | n |
| TCEI1. Treated with respect | 4.61 | 0.31 | 27 | 4.76 | 0.16 | 11 | -0.35, 0.06 | -1.48 | 36 | .07 |
| TCEI3. Usefulness of outside assignments | 4.07 | 0.51 | 25 | 4.34 | 0.30 | 11 | -0.61, 0.07 | -1.63 | 34 | .11 |
| TCEI5. Usefulness of online tools and technology | 4.19 | 0.48 | 25 | 4.39 | 0.38 | 11 | -0.53, 0.13 | -1.23 | 34 | .44 |
| TCEI6. Course difficulty | 3.29 | 0.31 | 25 | 3.26 | 0.27 | 11 | -0.19, 0.26 | 0.29 | 34 | .33 |
| TCEI8. Work required | 3.58 | 0.28 | 25 | 3.48 | 0.34 | 11 | -0.12, 0.32 | 0.93 | 34 | .92 |
| TCEI9. Previous learning | 3.41 | 0.37 | 25 | 3.29 | 0.29 | 11 | -0.14, 0.37 | 0.89 | 34 | .41 |
| TCEI10. Effort | 3.56 | 0.31 | 25 | 3.57 | 0.28 | 11 | -0.23, 0.21 | -0.10 | 34 | .97 |
| TCEI12. Active Learning Strategies | 4.27 | 0.48 | 25 | 4.39 | 0.30 | 11 | -0.44, 0.20 | -0.77 | 34 | .42 |
| TCEI13. Inspires Interest | 4.13 | 0.57 | 25 | 4.44 | 0.27 | 11 | -0.68, 0.05 | -1.73 | 34 | .06 |
| TCEI14. Amount online activities | 3.38 | 0.55 | 20 | 3.14 | 0.86 | 8 | -0.31, 0.80 | 0.91 | 26 | .28 |
| TCEI15. Hours per week | 3.98 | 1.34 | 25 | 3.66 | 0.62 | 11 | -0.55, 1.18 | 0.74 | 34 | .08 |
| TCEI16. Valuable hours | 3.91 | 0.54 | 25 | 4.20 | 0.28 | 11 | -0.64, 0.06 | -1.67 | 34 | .16 |
| TCEI17. Expected grade | 6.51 | 0.26 | 27 | 6.48 | 0.16 | 11 | -0.14, 0.21 | 0.40 | 36 | .18 |
| TCEI19. Class level | 2.86 | 1.12 | 27 | 2.56 | 0.65 | 11 | -0.44, 1.04 | 0.83 | 36 | .09 |
| TCEI20. Course program fulfillment | 2.52 | 0.51 | 27 | 2.56 | 0.45 | 11 | -0.40, 0.32 | -0.24 | 36 | .78 |
| TCEI22. Overall teaching effectiveness | 4.10 | 0.62 | 27 | 4.40 | 0.34 | 11 | -0.70, 0.11 | -1.48 | 36 | .12 |
| TCEI24. Overall course rating | 3.77 | 0.61 | 27 | 4.11 | 0.40 | 11 | -0.74, 0.08 | -1.63 | 36 | .13 |

**Instructor Feedback**

Seven instructors provided qualitative feedback regarding the difference in information gathered by the pilot TCE and the current TCE items. Four felt the pilot items provided a report that was clear and helpful/useful for understanding the student experience. One instructor mentioned the pilot items seemed to elicit more thoughtful responses from students, and provided a clearer picture on what students learned and their growth in the course. Another felt they were able to get clear take-aways from the student experience.

The criticisms regarding the pilot items mentioned: no item directly about the instructor, some items seem redundant, and did not ask about overall effectiveness or instructor comparison. One instructor felt the item regarding learning objectives may be outdated and should be changed to reflect competencies instead.

**Limitations**

*Response Rate and Students Enrolled*

There was a wide range of response rates (0.03-0.88) with an average of 0.43 and a median of 0.39. The validity of the TCEs increases with the response rate. Nulty (2008) a minimum response rate of approximately 0.58 is needed for validity depending on class size, with smaller classes (up to 20 students) requiring at least 58% and larger courses as low as 35%.

There was also a lot of variance in the number of students enrolled, with a range of 5 to 791. The mean students enrolled was 90, with a median of 45.00. The range of students enrolled reduces the statistical power of the analysis.

*Small Sample Size*

Based on an a priori power analysis, to achieve power of 0.95 a sample size of 176 total courses, with gender groups of 88 each, is required. With only 39 courses and prospective TCE surveys, this is a small sample of the total courses receiving the current TCE surveys. This reduces the power of the analysis and leaves it vulnerable to Type II errors.

*Uneven Group Size*

Instructor gender and course mode of delivery groups are not even, which should be considered when using a t-test. The female-instructor group was more than twice as large as the male-instructor group. Since only the female group could be randomized and the sample size was so low, the analysis was run with all possible samples. When a larger data set is obtained, the group sizes should be equalized and randomized.

**Future Directions**

If these questions are selected to replace the current TCE, another analysis should be completed with at least 176 course samples. Further analyses to investigate differences in gender, colleges, course format, and other course attributes should be completed with the larger data set as well.

**References**

Clayson, D. E. (2013). Initial impressions and the student evaluation of teaching. Journal of Education for Business 88:26-35.

Hamermesh, D. S. and Parker, A. M. (2003). Beauty in the classroom: Professorial pulchritude and putative pedagogical productivity. Economics of Education Review 24:369-376.

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(4):291-303.

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: what can be done?. Assessment & Evaluation in Higher Education, 33(3), 301-314.

Spooren, P, Brockx, B, Mortelmans, D. (2013). On the validity of student evaluation of teaching: the state of the art. Review of Educational Research 83(4):598-642.

Stark, P. B. and Freishtat, R. An evaluation of course evaluations. Science Open Research, DOI: [10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1](http://dx.doi.org/10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1).