Journal of Collective Bargaining in the Academy

Volume 0 *National Center Proceedings 2015* Article 48

April 2015

Teaching evaluations: class act or class action?

Philip B. Stark
UC Berkeley

Teaching evaluations: class act or class action?

Philip B. Stark
Department of Statistics
University of California, Berkeley

National Center for the Study of Collective Bargaining in Higher Education and the Professions

Annual Conference

Hunter College

New York, NY
19–20 April 2015

1 / 19

What do SET measure? No consensus.

- SET scores are highly correlated with students' grade expectations
  Marsh & Cooper, 1980; Short et al., 2012; Worthington, 2002

What do SET measure? No consensus.

- SET scores are highly correlated with students' grade expectations
  Marsh & Cooper, 1980; Short et al., 2012; Worthington, 2002
- SET scores & enjoyment scores *very* strongly correlated
  Stark, unpublished, 2014

What do SET measure? No consensus.

- SET scores are highly correlated with students' grade expectations
  Marsh & Cooper, 1980; Short et al., 2012; Worthington, 2002
- SET scores & enjoyment scores *very* strongly correlated
  Stark, unpublished, 2014
- SET can be predicted from the students' reaction to 30 seconds of silent video of the instructor; physical attractiveness matters
  Ambady & Rosenthal, 1993

What do SET measure? No consensus.

- SET scores are highly correlated with students' grade expectations
  Marsh & Cooper, 1980; Short et al., 2012; Worthington, 2002
- SET scores & enjoyment scores *very* strongly correlated
  Stark, unpublished, 2014
- SET can be predicted from the students' reaction to 30 seconds of silent video of the instructor; physical attractiveness matters
  Ambady & Rosenthal, 1993
- gender, ethnicity, & the instructor's age matter
  Anderson & Miller, 1997; Basow, 1995; Boring, 2014; Cramer & Alexitch, 2000; Marsh & Dunkin, 1992; McNell et al., 2014; Wachtel, 1998; Weinberg et al., 2007; Worthington, 2002

5 / 19

What do SET measure? No consensus.

- SET scores are highly correlated with students' grade expectations
  Marsh & Cooper, 1980; Short et al., 2012; Worthington, 2002
- SET scores & enjoyment scores *very* strongly correlated
  Stark, unpublished, 2014
- SET can be predicted from the students' reaction to 30 seconds of silent video of the instructor; physical attractiveness matters
  Ambady & Rosenthal, 1993
- gender, ethnicity, & the instructor's age matter
  Anderson & Miller, 1997; Basow, 1995; Boring, 2014; Cramer & Alexitch, 2000; Marsh & Dunkin, 1992; McNell et al., 2014; Wachtel, 1998; Weinberg et al., 2007; Worthington, 2002
- omnibus questions about curriculum design, effectiveness, etc., appear most influenced by factors unrelated to learning
  Worthington, 2002

The gold standard: Randomized, controlled experiments

The gold standard: Randomized, controlled experiments

Carrell & West, 2008

United States Air Force Academy assigns students to instructors at random in core courses, including follow-on courses. All sections have identical syllabi and exams.

Student evaluations are positively correlated with contemporaneous professor value-added and negatively correlated with follow-on student achievement.

That is, students appear to reward higher grades in the introductory course but punish professors who increase deep learning (introductory course professor value-added in follow-on courses).

The gold standard: Randomized, controlled experiments

Carrell & West, 2008

United States Air Force Academy assigns students to instructors at random in core courses, including follow-on courses. All sections have identical syllabi and exams.

Student evaluations are positively correlated with contemporaneous professor value-added and negatively correlated with follow-on student achievement.

That is, students appear to reward higher grades in the introductory course but punish professors who increase deep learning (introductory course professor value-added in follow-on courses).

Braga, Paccagnella, & Pellizzari, 2011

Randomized assignment of students to instructors at Bocconi University, Milan

in other words, teachers who are associated with better subsequent performance receive worst evaluations from their students.

9 / 19

McNell, Driscoll & Hunt, 2014: Gender Bias

NC State online course.

Randomized assignments of students into 4 groups.

2 instructors, 1 male 1 female.

Each instructor was identified to students by actual gender in 1 section, false gender in 1 section.

Regardless of actual gender, substantially higher ratings when each instructor was identified as male, even for "objective" measures, e.g., speed of returning homework.

5-point scale.

Adjective	F - M
Caring	-0.47
Consistent	-0.57
Enthusiastic	-0.76
Fair	-0.47
Feedback	-0.46
Helpful	-0.35
Knowledgeable	-0.67
Praise	-0.61
Professional	-0.80
Prompt	-0.61
Respectful	-0.22
Responsive	-0.61

Boring, 2014: more evidence of gender bias

> Male students in particular tend to give higher overall satisfaction scores to male teachers, rewarding them for their perceived higher quality in course delivery style. … Male teachers can increase their SET scores by investing more effort in the characteristics that male students tend to value more. However, female teachers must invest more effort improving the teaching dimensions in which students tend to perceive a slight comparative advantage for women, i.e. course structure, organization and teaching material.…
>
> The results suggest that better teaching is not necessarily measured by SETs.

Lauer, 2012: Student comments knotty, too

Survey of 185 students, 45 faculty at Rollins College, Winter Park, Florida

I once believed that narrative comments on course evaluation forms were straightforward and useful.

Lauer, 2012: Student comments knotty, too

Survey of 185 students, 45 faculty at Rollins College, Winter Park, Florida

> I once believed that narrative comments on course evaluation forms were straightforward and useful.

Faculty & students ascribe quite different meanings to words such as "fair," "professional," "organized," "challenging," & "respectful."

Lauer, 2012: Student comments knotty, too

Survey of 185 students, 45 faculty at Rollins College, Winter Park, Florida

> I once believed that narrative comments on course evaluation forms were straightforward and useful.

Faculty & students ascribe quite different meanings to words such as "fair," "professional," "organized," "challenging," & "respectful."

***not fair* means …**	**student %**	**instructor %**
plays favorites	45.8	31.7
grading problematic	2.3	49.2
work is too hard	12.7	0
won't "work with you" on problems	12.3	0
other	6.9	19

Benton & Cashin, 2012: exemplar SET apologists

> It is difficult to get a man to understand something, when his salary depends upon his not understanding it! —Upton Sinclair

- Widely cited, unrefereed technical report from a business that sells SET; flawed statistics

Benton & Cashin, 2012: exemplar SET apologists

> It is difficult to get a man to understand something, when his salary depends upon his not understanding it! —Upton Sinclair

- Widely cited, unrefereed technical report from a business that sells SET; flawed statistics
- Rebut straw man positions:
  - Students cannot make consistent judgments.
  - Student ratings are just popularity contests.
  - Students will not appreciate good teaching until they are out of college a few years.
  - Students just want easy courses.
  - Student feedback cannot be used to help improve instruction.

16 / 19

Benton & Cashin, 2012: exemplar SET apologists

> It is difficult to get a man to understand something, when his salary depends upon his not understanding it! —Upton Sinclair

- Widely cited, unrefereed technical report from a business that sells SET; flawed statistics
- Rebut straw man positions:
  - Students cannot make consistent judgments.
  - Student ratings are just popularity contests.
  - Students will not appreciate good teaching until they are out of college a few years.
  - Students just want easy courses.
  - Student feedback cannot be used to help improve instruction.
- The two non-absolutist statements they reject are demonstrably true:
  - Student ratings are unreliable and invalid.
  - The time of day the course is offered affects ratings.

Benton & Cashin, 2012: exemplar SET apologists

> It is difficult to get a man to understand something, when his salary depends upon his not understanding it! —Upton Sinclair

- Widely cited, unrefereed technical report from a business that sells SET; flawed statistics
- Rebut straw man positions:
  - Students cannot make consistent judgments.
  - Student ratings are just popularity contests.
  - Students will not appreciate good teaching until they are out of college a few years.
  - Students just want easy courses.
  - Student feedback cannot be used to help improve instruction.
- The two non-absolutist statements they reject are demonstrably true:
  - Student ratings are unreliable and invalid.
  - The time of day the course is offered affects ratings.
- The remaining statement they reject is true, in my experience as a teacher and department chair:
  - Emphasis on student ratings has led to grade inflation.

  See also Ewing, 2012; Isely & Singh, 2005; Krautmann & Sander, 1999; McPherson, 2006

18 / 19

Recommendations

1. Drop omnibus items about "overall teaching effectiveness" and "value of the course"
2. Do not average or compare averages of SET scores: Such averages do not make sense statistically. Instead, report the distribution of scores, the number of responders, and the response rate.
3. Responders are not a random sample and there's no reason their responses should be representative of the class as a whole: do not extrapolate.
4. Pay attention to student comments but understand their limitations and heed differences in language usage.
5. Avoid comparing teaching effectiveness across courses of different types, levels, sizes, functions, or disciplines.
6. Use teaching portfolios as part of the review process.
7. Use classroom observation as part of milestone reviews.
8. To improve teaching and evaluate teaching fairly and honestly, spend time observing teaching & teaching materials.