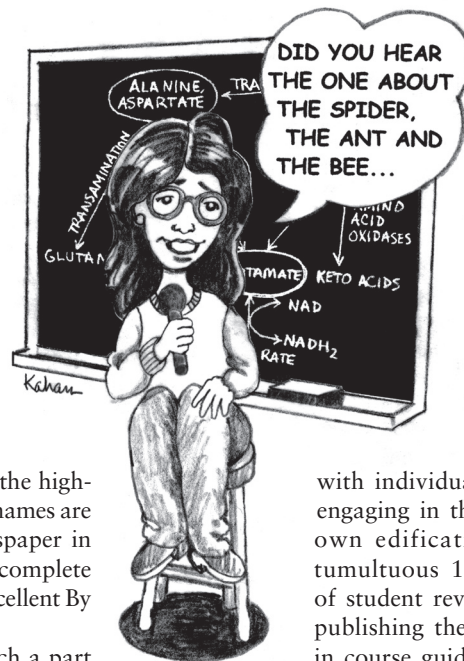


Incomplete Teaching

May Berenbaum



I like to think that I'm a pretty good teacher, but there have been many moments over the 26 years I've been teaching that I've had grave doubts. I devote an inordinate amount of time to my classes every semester. But that I do so is not so much a reflection of my love of teaching as it is of a paralyzing fear of standing up unprepared in front of a couple of hundred students who signed up for the class I teach only because a careful examination of the course catalogue led to the realization that there was no other practical way of satisfying the campus general education requirement for 3 hours of biology (a requirement that they resent deeply).

And even for the general education students who actually enjoy the class, I sometimes wonder how effective I am in inculcating the fundamentals of entomology. Once, while I was in Campustown with my daughter during a street fair, we stopped in front of a coffee shop where free samples of Italian ices were being handed out. When the barista saw me, she exclaimed, "Professor Berenbaum! I took Ent. 105 last spring and I loved it! It was my favorite course!" I thanked her and smiled proudly at my daughter, right up to the moment when the barista waved away some yellowjackets buzzing around the free samples and exclaimed, "I wish these bees would leave me alone!" Fortunately, my daughter, who learned the difference between yellowjackets and bees back in first grade, was too busy with her tangerine Italian ice to snigger at me out loud.

Part of the reason for my lingering doubts is that there are very few accurate metrics for measuring teaching quality. On our campus, we rely on a survey, conducted at the end of every semester, which involves handing out Scantron forms (or ICES forms, for "Instructor and Course Evaluation System"). Students fill them out and send them to a campus office, which analyzes the results sta-

tistically and determines the highest-scoring faculty. Their names are listed in the school newspaper in what is known as the "Incomplete List of Teachers Rated Excellent By Their Students."

This process is as much a part of the cycle of life as the spring and fall migratory flights of birds, albeit with far greater potential impact on the ego. Even though it's an integral part of every faculty member's life, like those migratory bird flights, there's still a lot of mystery surrounding the Incomplete List. When, as an assistant professor, I first heard about the Incomplete List, I was immediately apprehensive. Why was this list incomplete? What was missing, and why? Was it just innocent oversight—say, some instructors may be rated excellent by their students but the forms were lost in campus mail? Or is it capricious, with the campus periodically dropping faculty whose names start with a vowel? Or is it more sinister, with faculty sometimes mysteriously disappearing after being rated excellent and thereby missing the list? After 26 years, I'm still not sure why the list is "incomplete," but the most mysterious aspect of the Incomplete List, and ICES forms in general (other than what the acronym actually stands for), is whether or not the forms reflect teaching excellence.

As it turns out, I'm not the only one who wonders—there's an extensive literature evaluating the evaluation of teaching and absolutely no consensus on what constitutes excellent teaching. In an effort to understand what teaching evaluations actually mean, I plunged into this literature to see what I could learn (in the process consuming hours I probably should have spent preparing lectures).

Universities didn't always control the process of course evaluation. Through the first half of the 20th century, it was a hit-or-miss, primarily instructor-driven affair,

with individual faculty members engaging in the process for their own edification. Then, in the tumultuous 1960s, in the spirit of student revolt, students began publishing their own evaluations in course guides often filled with what might best be characterized as, even if accurate, uncharitable

descriptions of courses and instructors. By the 1970s, universities basically took the process over. In doing so, they generated a tremendous and ongoing debate in the educational literature about the validity and fairness of the process.

It was about this time that the most famous publication about teaching evaluation ever written appeared—"The Doctor Fox Lecture: A paradigm of educational seduction" (Naftulin et al. 1973). These authors hired a professional actor who "looked distinguished and sounded authoritative," to portray an "authority on the application of mathematics to human behavior." They gave him the plausible name of Dr. Myron L. Fox, equipped him with an entirely fabricated but very "impressive" vita and managed to get him invited to lecture to three groups of professional educators. His topic was "Mathematical Game Theory as Applied to Physician Education;" and he was provided with a lecture text characterized by deliberate "excessive use of double talk, neologisms, non sequiturs, and contradictory statements. All this was to be interspersed with parenthetical humor and meaningless references to unrelated topics."

Which reminds me—did you hear about the termite who walked into a tavern and asked the first customer he met, "Is the bartender here?" I love that joke—as insect jokes go, that one kills.

Well, invariably, people loved Dr. Fox—70, 82, and 90% felt that he "presented his material in a well organized format"; 80 to

90% felt that he “put his material across in an interesting way”; and one person even claimed to have read his earlier papers on the subject. Between 87 and 100% of the audience said that they were “stimulated by the lecture.” These results led the authors to identify what they called “educational seduction” also known as the “Dr. Fox effect”—an illusion of having learned something “if the lecturer simulates a style of authority and wit.”

A propos of nothing, and in keeping with the tradition of non sequitur and meaningless references to unrelated topics, in the course of searching for Dr. Fox, I also found the “Aunt Fanny effect” in teaching evaluation. Also known as the Barnum effect, it’s the “consequence of one’s belief that a vague personality description truly describes oneself, when in reality that description may apply to almost anyone; sometimes referred to as the “Aunt Fanny effect” because the same personality might be applied to anyone’s Aunt Fanny” (Cohen and Swerdlik, 2002).

The Dr. Fox study has since been criticized for its methodology. In fact, problems arose in the 1970s with any experimental tests of validity due to ethical concerns. In retrospect, it’s remarkable that these experiments were ever done: “Grade manipulations imposed stresses and used deceptions that university human subject review committees do not look too kindly upon” (according to Williams and Ceci 1997).

As a consequence, in the 1980s, investigators chose an alternative path, and dozens, if not hundreds, of correlational and theoretical analyses were published purporting to test validity via aggregate statistical summaries. By 1988, Cashin reported the existence of more than 1,300 publications on instructional evaluation; Feldman (1989) pegged the number at more than 2,000.

Subsequent work has done little to provide guidance on improving effectiveness. Ambady and Rosenthal (1993) conducted a study with Harvard students who were shown content-free video clips, 30 seconds long and without audio, of graduate teaching fellows and then asked to evaluate these individuals. At the end of the semester, the student evaluations for these graduate teaching fellows were compared with the evaluations based on the 30-second content-free video clips and were found to be significantly correlated, with an r value of 0.76.

Knowing that first 30 seconds of class may determine whether or not I’ll make it onto the Incomplete List at the end of the semester has made me even more worried about going to class with spinach stuck between my teeth or (as actually happened not too long ago) with cupcake icing on my sleeve.

Then there’s the study of Williams and

Ceci (1997); these Cornell professors arranged to teach the same undergraduate developmental psychology class for two consecutive semesters exactly identically, except that, during the second semester, the lecturer (Steven Ceci) used a “more enthusiastic” lecturing style (i.e., with more voice modulation and gestures). Content was identical, even down to the actual words (he had taught this class every year since 1977). The mean rating on “how knowledgeable is the instructor” rose from 3.61 fall to 4.05 spring; “how accessible was the instructor” rose from 2.00 to 4.06, “how organized” from 3.18 to 4.09, and the overall rating rose from 3.09 to 3.92, an increase of a full standard deviation and a difference significant at the level of $p < 0.0001$. “How much did you learn?” went from 2.93 to 4.05. Even the ratings for the textbook rose from 2.06 to 2.98 (rising from “poor” to “average”).

Admittedly, this study was criticized, too. D’Apollonia and Abrami (1997), responsible for earlier meta-analytical approaches, were harsh in their criticism; Williams and Ceci (1997b) themselves in a response wrote “—D’Apollonia and Abrami’s critique can be summarized as follows: ‘Our study is, in part, wrong; the part that’s not wrong is trivial; and the part that is neither wrong nor trivial they thought of first.’”

If even the professionals can’t decide whether the ways we have to evaluate teaching are valid or fair, then it’s not clear what I should do at this point. I’m extraordinarily reluctant to read the 2,000+ publications to decide if the latest ICES scores really reflect my value as a human being, at least in part because I have no idea what “construct validity theory” or the “Biglan model of faculty subcultures” is and finding out would mean having even less time to get lectures ready for the rest of the semester.

Actually, the Biglan model of faculty subcultures has something to do with differences among academic disciplines, and whether or not teaching entomology is more or less likely to earn a faculty member decent ratings isn’t readily discernible. Within the vast literature on teaching effectiveness and student ratings, entomology is conspicuous by its absence. In fact, I found only one article that mentions entomology even in passing. Hoyt and Reed (1977) evaluated student ratings of 266 faculty members at Kansas State University, including entomology faculty. They found that there was a “modest but significant correlation between ratings of teaching effectiveness and percent salary increase” (which raises the possibility for interesting negotiations between department heads and faculty with low teaching scores over pay raises).

Despite the essential absence of entomological pedagogy, I did stumble across an

article that might be useful after all. Based on a survey of 453 undergraduate students, Adamson et al. (2005), found that one attribute found to be “significantly related” to teaching effectiveness was “how funny the instructor was perceived.”

I’m not sure how applicable the finding might be to entomology, but it does remind me of the race between two silkworms. Have you heard about it? It ended in a tie.

References Cited

- Ambady, N., and R. Rosenthal. 1993. Half a minute: predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *J. Pers. Soc. Psychol.* 63: 431–441.
- Adamson, G., D. O’Kane, and M. Shevlin. 2005. Students’ ratings of teaching effectiveness: A laughing matter? *Psychol. Rep.* 96: 225–226.
- Cashin, W. E. 1988. Student ratings of teaching: a summary of the research. Center for Faculty Evaluation and Development, Manhattan, KS.
- Cashin, W. E. 1990. Students do rate different academic fields differently. In M. Theall and J. Franklin [Eds.]. *Student ratings of instruction: issues for improving practice. new directions for teaching and learning*, no. 43. Jossey-Bass, San Francisco.
- Cohen, R. J., and M. Swerdlik. 2002. *Psychological testing and assessment: an introduction to tests and measurement*, 5th ed. McGraw-Hill, New York. http://highered.mcgraw-hill.com/sites/0767421574/student_view0/chapter13/glossary.html
- d’Apollonia, S., and P. C. Abrami. 1997. In response... *Change* 29: 199–200.
- Feldman, K. A. 1989. The association between student ratings of specific instructional dimensions and student achievement: refining and extending the synthesis of data from multisection validity studies. *Res. Higher Ed.* 30: 583–645.
- Naftulin, D. H., J. E. Ware, Jr., and F. A. Donnelly. 1973. The Doctor Fox lecture: a paradigm of educational seduction. *J. Med. Ed.* 48: 630–635.
- Neumann, Y., and L. Neumann, 1982. Characteristics of academic areas and students’ evaluation of instruction. *Res. Higher Ed.* 19: 323–334.
- Williams, W. M., and S. J. Ceci. 1997a. “How’m I doing?” Problems with student ratings of instructors and courses. *Change* 29: 12–23.
- Williams, W. M., and S. J. Ceci. 1997b. The authors respond. *Change* 29: 19.



May Berenbaum is a professor and head of the Department of Entomology, University of Illinois, 320 Morrill Hall, 505 South Goodwin Avenue, Urbana, IL 61801. Currently, she is studying the chemical aspects of interaction between herbivorous insects and their hosts.

