Studies in Educational Evaluation 58 (2018) 60–69

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

Evaluating the quality of middle school mathematics teachers, using videos rated by college students

Gerhard Sonnert[a,*], Zahra Hazari[b], Philip M. Sadler[c]

[a] Harvard-Smithsonian Center for Astrophysics, United States
[b] Florida International University, United States
[c] Harvard-Smithsonian Center for Astrophysics, United States

ARTICLE INFO

Keywords:
Mathematics teachers
Middle school
Video evaluation

ABSTRACT

Using experts in teaching to evaluate pre-college mathematics teachers is both time consuming and costly. This study examines the potential of letting undergraduate mathematics students perform this task, comparing their ratings of two dimensions, richness of instruction and mathematical correctness, to those previously assigned by an expert. Using 85 undergraduates of two U.S. institutions, who independently watched short videos of teachers, we found that student ratings of teachers' correctness were a good match to the expert ratings; student ratings of richness of instruction, less so. A "halo effect" was observed in that students did not fully differentiate between richness and correctness in their ratings. Moreover, male students gave harsher ratings than did female students. Whereas undergraduates showed promise in accurately evaluating teachers of younger students, improvements in terms of rating items, attention to bias, and explicit training on the teaching dimensions to be rated should be addressed.

1. Introduction

The assessment of teachers is essential to the process of improving the quality of instruction. Whereas survey-based evaluations of instructors by students are near universal in higher education (Linse, 2017; Onwuegbuzie, Daniel, & Collins, 2009), the most common method of evaluating teachers at the high school level and below is practice-based teacher evaluation. Often, this is done through classroom observation by principals or similar senior personnel. Even though this is a "time-intensive process (Tyler, 2011, p. 6)," as well as "relatively costly (Tyler, 2011, p. 6)," it is not unproblematic. It was, for instance, found riddled with leniency bias (e.g., Tucker, 1997; Weisberg, Sexton, Mulhern, & Keeling, 2009) and of uneven quality (Bergin, Wind, Grajeda, & Tsai, 2017). In addition, these types of evaluations provide a limited perspective because they come from a small number of individuals whose visits are often pre-arranged, reducing the validity of capturing a teacher's typical practice. Another persistent problem that has been found in teacher evaluations has been gender bias.

Hence, there is a need to find new ways to evaluate teachers. In the U.S., a massive effort to determine teacher effectiveness has been carried out in the "Measures of Effective Teaching" (MET) project supported by the Gates Foundation (Ho & Kane, 2013; Kane & Staiger, 2012; Kane, McCaffrey, Miller, & Staiger, 2013; cf. Cohen & Goldhaber, 2016). In many other countries, too, increasing emphasis has been placed on teachers' performance evaluation (e.g., Flores, 2012; Liu & Zhao, 2013; Range, Scherz, Holt, & Young, 2011) in the common belief that this is a key element in improving teacher quality and, consequently, student learning. Several sophisticated protocols for teacher evaluation have been developed, such as the Classroom Assessment Scoring System (CLASS), the Mathematical Quality of Instruction (MQI), the Protocol for Language Arts Teacher Observations (PLATO), and the UTeach Teacher Observation Protocol (UTOP) (Kane & Staiger, 2012), and evaluation-based reforms have been tried out in school systems like Washington, DC (Dee & Wyckoff, 2015), and Chicago (Steinberg & Sartain, 2015). Yet it has remained rather unclear if current systems of teacher evaluation have the hoped-for beneficial effects (Hallinger, Heck, & Murphy, 2014). As one of the potential directions of moving forward, the *Designing High Quality Evaluation Systems for High School Teachers* report by the Center for American Progress advocates "Exploring, developing, and testing the increased use of technology such as classroom video recording as a means for generating efficiency and productivity gains in practice-based evaluation (Tyler, 2011, p. 4)."

This study explores the feasibility and promise of evaluations of brief video-recorded classroom scenes showing U.S. middle school mathematics teachers at work. It takes the novel approach of using

* Corresponding author.
E-mail address: gsonnert@cfa.harvard.edu (G. Sonnert).

https://doi.org/10.1016/j.stueduc.2018.05.006
Received 1 December 2017; Received in revised form 5 April 2018; Accepted 11 May 2018

college mathematics students to watch and evaluate teachers. The proposition to be tested is that these students, who several years prior took middle school mathematics, have the perspective and knowledge to make ratings that are similar to those of an expert on the teaching of middle school mathematics.

2. Literature review

The validity of ratings of brief video clips rests on the premise that observers are able to make reasonable judgments in a short time frame. Social psychologists have studied humans' amazing tendency to make rapid judgments about other humans (Ambady & Rosenthal, 1992), sometimes in as little as 100 ms (Willis & Todorov, 2006). Whereas many of these studies have been done in a variety of contexts quite different from teacher evaluation, Ambady and Rosenthal (1993) had undergraduate students rate college teachers' nonverbal behavior traits from very brief (under 30 s) silent video clips (cf. Babad, Avni-Babad, & Rosenthal, 2004; Tom, Tong, & Hesse, 2010). Surprisingly, these ratings were found to predict end-of-semester student evaluations of these teachers by their undergraduate students. In a second experiment, which is even more relevant to our topic, Ambady and Rosenthal (1993) repeated the investigation with high school teachers. Undergraduate student ratings predicted the principal's rating of the teachers' overall effectiveness. The student raters were also found to exhibit a high degree of consensus.

Whereas these results justify some optimism about the viability of teacher evaluations using short video clips, there are major concerns, among them: First, the evaluations might end up focused on irrelevant teacher characteristics. For instance, ratings might be biased in favor or against certain groups (and, in this article, we focus particularly on a potential gender bias). Then, there might be over-generalization—a general opinion formed about the teacher might overwhelm performance differences in different categories, i.e., ratings across categories may be highly correlated. These possible pitfalls as well as the potential strengths noted above motivate further examination.

In general, students' evaluation of their teachers has long been a focus of intense study (e.g., Marsh, 1984; Marsh & Roche, 1997; for an overview, see Benton & Cashin, 2012). Scherer, Nilsen, and Jansen (2016) found that students' ratings of instructional quality had three underlying dimensions (teacher support, cognitive activation, and classroom management), and Kunter and Baumert (2006) also reported three evaluative dimensions (classroom management, cognitive autonomy, and tempo of interaction), which were shared by students and teachers. Importantly, students' ratings have been consistently found to be influenced by the teachers' personality traits. Harking back to a sequence of famous early studies, this influence is sometimes called the "Dr. Fox effect" (Naftulin, Ware, & Donnelly, 1973; Ware & Williams, 1975). When a lecture of rather meaningless information content was presented by an actor portraying "Dr. Myron L. Fox" in a charismatic manner, the audience gave high satisfaction ratings. (This effect was not only observed with students, but also with professional educators.) The Dr. Fox effect can be considered a particular instance of the more general "halo effect." Long known in social psychology, the halo effect describes the tendency to let one's evaluation of a person's particular trait color one's evaluation of that person's other traits (Thorndike, 1920; cf. Nisbett & Wilson, 1977; Wade & DiMaria, 2003). Similarly, if raters have a generally favorable impression of a person, they tend to give that person high evaluations on their specific traits. Whereas, in this case, the halo of general goodness makes the person look better in all specific aspects, the opposite—a negative halo effect, so to speak–of course can also operate. Showing a kind of temporal halo effect, a study of teachers and administrators rating videos of other teachers found that first impressions about the ratees' performance tended to linger and shade subsequent ratings (Ho & Kane, 2013). While we are concerned with the halo effect in instructors' evaluations by students, we should add that a halo effect in the reverse direction (in students' evaluations by teachers) has also been documented (Abikoff, Courtney, Pelham, & Koplewicz, 1993; Foster & Ysseldyke, 1976; cf. Meissel, Meyer, Yao, & Rubie-Davies, 2017).

A study by Radmacher and Martin (2001) found teacher extraversion as the sole significant predictor of college student evaluations of teacher effectiveness, after controlling for other factors. In a study of undergraduate psychology and business and management courses, in which Myers-Briggs Type Indicators were obtained for both instructors and students, extraverted teachers were rated higher in effectiveness than were introverted teachers, intuitive teachers higher than sensing ones, and feeling teachers higher than thinking teachers, regardless of the student raters' personality (Hart & Driver, 1978). In addition, feeling students gave higher scores than thinking students. Focusing on the Big Five personality traits of instructors in seven general education courses at a small Midwestern university, Patrick (2011) found that student ratings of their instructors' extraversion, openness, agreeableness and conscientiousness positively correlated with their ratings of instructor effectiveness. Other studies at the college level similarly identified the teacher's perceived expressiveness, warmth and enthusiasm as factors that positively affected students' evaluations of their teacher (Basow & Distenfeld, 1985; Basow & Silberg, 1987). Even an instructor's physical attractiveness was found to boost college students' ratings of instructor warmth, sensitivity, ability to communicate, knowledge of subject matter, and superiority (Lombardo & Tocci, 1979; cf. Dion, Berscheid, & Walster, 1972; Landy & Sigall, 1974).

Another focus of research has been to examine the potential gender bias in ratings of teachers. The results of prior studies in this area have been mixed. Haemmerlie and Highfill (1991) found no gender bias in the ratings of male and female instructors by undergraduate males majoring in technical engineering fields. In non-science introductory courses, Bennett (1982) found no direct anti-female bias in students' ratings of their professors. On the contrary, female instructors were judged to be warmer and more supportive than male instructors, and this accounted for women instructors' higher ratings on overall effectiveness, as well as other aspects of teaching performance. However, students also expected greater interpersonal support from female professors than from male professors.

In Kaschak's (1978) study, female college students showed no difference in their ratings, on various scales, of female and male instructors, but the male students viewed the male instructors more favorably than they did the female instructors. Similarly, Basow and Silberg (1987) found that, while female college students gave similar evaluations to their male and female professors, the male students gave higher ratings to the male professors than to the female professors. Another interaction between student gender and teacher gender was observed by Basow (1995) who found that the student ratings of the male professors at a small liberal arts college did not differ by the students' gender, whereas female professors received higher ratings from female than from male students. Martin (1984) found that students tended to rate instructors of the same sex higher than instructors of the opposite sex. In another study, when college students were asked to describe the "best" professor they had had in college, that professor was female more often for female students than for male students (Basow, 2000). Using an online class as their experimental site, MacNell, Driscoll, and Hunt (2015) were able to assign the same instructors two different gender identities under which they taught different sections of the course. This study, which allowed a clear view on the effect of perceived instructor gender, detected a gender bias in students' ratings in favor of (perceived) male instructors. Arbuckle and Williams (2003) found an interaction effect of teachers' gender and age. In their experiment, college students gave higher ratings on speaking enthusiastically to professors who were described as young and male.

Desirable teacher attributes were often defined differently for male and female teachers and according to gender stereotypes, in essence creating a double standard (Basow & Silberg, 1987; Kierstead, D'Agostino, & Dill, 1988; Martin, 1984). Basow (1990) found that the