3-2022

The Biasing Impact of Positive Instructor Reputation on Student Evaluations of Teaching

D. Brian McNatt
Boise State University

The Biasing Impact of Positive Instructor Reputation on Student Evaluations of Teaching

D. Brian McNatt*
Boise State University
brianmcnatt@boisestate.edu

Abstract

A naturally-occurring intervention in a longitudinal field setting (4 months) was used to examine the presence and biasing impact of a positive reputation on subsequent ratings of work performance (student evaluations of teaching). During pre-semester interactions, first-year MBA students received information from second-year MBAs about their upcoming professors and classes. Favorable information about the two professors and course examined in the present study caused a positive reputation. Results indicated that despite four months of experiencing actual performance, the positive reputation hindered students' decision-making process resulting in biasedly inflated ratings of instructor performance and halo error judgments of course materials, grading, and amount learned. The problematic implications of using biased student evaluations of teaching to measure faculty performance is discussed, along with suggestions of ways to mitigate against overreliance on this evaluation method and to possibly minimize reputational effects.

Keywords: reputation, student evaluations of teaching, perception bias, belief perseverance, performance appraisal

1. Introduction

Years ago, as I completed my Ph.D. degree, I was interviewing at a prestigious university. The job search committee chair surprised me in his introduction to the department faculty, before my beginning my job talk. He shared that I was getting my degree from a top university, that even though I was only a Ph.D. student, that one of my publications was a solo-authored piece in the best journal in the field, and that I had worked with and been mentored by three of the top HR/OB scholars in the world. As I began my talk I was thinking, "Wow, I haven't even started and they already think I'm great."

When we classify a target a certain way, we form expectations that can impact our perceptions of what we then see and experience (Baltes & Parker, 2000b). Scholars have studied how such preconceived notions, labels, or expectations, can modify our perceptions and impact the accuracy of our judgment and decision making (e.g., DeNisi, Cafferty, & Meglino, 1984; Martell & Willis, 1993). Even erroneous information has been found to create initial impressions that can bias perceptions. Once formed, these views can become so entrenched that they may remain despite receiving contradictory evidence or being made aware of the deception (Misra, 1992; Turban, Forret, and Hendrickson, 1998). This can have serious ramifications when making important decisions.

A particularly important judgment within organizations are performance evaluations. They influence fundamental employment decisions such as hiring, promotions, and termination, and ultimately employee development and organizational effectiveness (Chow, 2004). Scholars have underscored the importance of accurate selection and performance decisions (Baltes & Parker, 2000a). However, if perceptual biases are present, they can reduce the accuracy and diminish the usefulness of such ratings (Balzer & Sulsky, 1992). Thus, researchers have emphasized the need to study the extent to which raters make evaluations based on previously formed judgments versus actual behaviors, the possible factors that create pre-conceived notions, and the impact they have on performance ratings (Baltes & Parker, 2000a, DeNisi et al., 1984).

The current study examines a potential bias that has received insufficient research attention, the ratee's reputation (Ferris et al., 2003; Towler & Dipboye, 2006), and an important performance evaluation within academia—student evaluations of teaching (SET). The vast majority of research studying reputation has been at the firm-level (versus

individual), with studies examining how organizations create goodwill or status in order to leverage and benefit from such positive social evaluations; as well as the detrimental effects of highly publicized incidents in which reputation is tarnished and stakeholder trust violated (George, et al., 2016).

Therefore, researchers have called for studies to investigate informal channels in which information spreads contagiously among people within an organization to create an individual's reputation (Dineen and Allen, 2016; George, et al., 2016). Additional calls for future research specifically highlight the need to explore the effects of reputation on job performance (ratings), promotions, and career success, and particularly using longitudinal research designs (Zinko, et al., 2007). In response, the present study examines reputation (1) at the individual level, (2) occurring from informal communications that naturally happen among people, and (3) how effects may persist over time and bias future evaluations even after direct experience with the individual.

2. Reputation and Ratings of Teaching Performance

A general definition of reputation is a collective social judgment, belief, or perceptions held regarding the quality or capabilities of a focal actor within a specific domain (George, et al., 2016). These impressions may be based on evaluations, rumors, and information passed on from others (Raub and Weesie, 1990). As such, reputation is a socially constructed concept, and can be partly understood through the lens of social contagion theory (Ferris, et al., 2003). Based on this theory, reputations are generated by relevant people who discuss observed or reported behaviors through informal conversations (gossip), such that the audience that is being informed of the reputation of another need not have direct contact with that individual (Carroll, et al., 2003; Zinko, et al., 2007).

Research has found various benefits that have accrued from having a positive reputation. For example, among people who displayed the same number of helpful behaviors, those with good reputations were perceived differently and received more rewards than the other helpful people (Johnson, et al., 2002); and athletes who worked with high-reputation coaches benefitted in the labor market (Kilduff, et al., 2016). Other studies have found an organization's positive reputation to be correlated with the ability to charge premium prices (Rindova, et al., 2005), higher starting salaries of business school's graduates (Boyd, Bergh, and Ketchen, Jr., 2010), and attracting more applicants, higher-quality applicants, and experiencing lower turnover (Dineen and Allen, 2016; Fuller and Delorey, 2016; Turban and Cable, 2003).

Such studies in the reputational literature may have an unspoken assumption—that the reputation is an accurate assessment of superior quality based on factual past behaviors; and thus, that others' actions in response to the reputation are wholly warranted (e.g., paying a huge premium or salary for a given product/employee). However, is it possible that believing the reputation of a given target, causes decision makers to biasedly exaggerate the deference or castigation due the target? For example, could actual positive circumstances warrant a 10% premium price for a product, but not the 25% premium charged by the organization and paid by many customers? This possibility of an exaggerated response/evaluation is explored in the present study.

Ratings of teaching performance are most often conceptualized as student evaluations of teaching—SET (Marsh, et al., 1997; Munz and Munz, 1997). In fact, SETs are often the primary, or perhaps exclusive measure of professors' teaching performance (Waldman, 2008)—although some progressive institutions have begun encouraging or using additional methods. The accuracy of SETs are thus paramount since they are used in decisions related to professors' contract renewal, raises, tenure, and promotion (Munz and Munz, 1997; Pan, et al., 2021). In business organizations, supervisors' ratings of others' performance has often been found to be influenced by things not reflective of actual job performance, and thus contain biases and distortions (e.g., Higgins, Judge, and Ferris, 2003). Similarly, academics have raised concerns with the potential biases in student evaluations (e.g., Bachen et al., 1999).

Some indicate that SETs encourage reducing student workloads and grade inflation, are not relevant across academic disciplines (Constanda, Clarke, and Morgan, 2018), and are primarily measures of professor and course liking (Pan, et al., 2021). Other academics assert that SETs are reliable and reasonably valid (e.g., Marsh, et al., 1997). However, these academics also accept that SETs could be biased if they were influenced by something unrelated to teaching effectiveness. This could be the case if professor ratings were biasedly impacted by an initial reputation held by students. This potential confound is more relevant now than ever before due to heightened competitiveness, the greater reliance on contractual instructors over tenured faculty (Hubbard, 2003), and the exponential increase in the ability to share information through technology. This can be seen in the proliferation of on-line professor ratings sites (e.g., ratemyprofessor.com) where students can post and read comments and evaluations of professors. Students might easily

develop an opinion (reputation) of their instructors prior to any interaction with them. Could this reputation then distort/bias the students' perceptions and future evaluation of those instructors? It is important to better understand this possibility.

Yet, a search of the SET literature reveals that there are relatively few empirical studies, and much of the experimental work conducted contains constrained designs. These include using laboratory settings with created scenarios, interventions using fabricated performance information to set students' expectations of fictitious instructors, such as a short (5-20 minute) video or a one-page written description of performance, and non-longitudinal designs (e.g., a 1-hour experimental session) (e.g., Johnson, et al., 2002; Towler and Dipboye, 2006). For this reason, the most consistent call for future studies is for field research, with real stimuli to create an actual reputation, and to examine any biased effect on performance appraisals (SETs) after exposure to repeated performance over a period of months (e.g., Baltes and Parker, 2000b; Sumer and Knight, 1996). To address these limitations, the present experiment is a (1) longitudinal, (2) field experiment, (3) with a naturally occurring positive reputation stimulus, that (4) examines the impact of the reputation on evaluations of instructor performance and other related items.

3. Theoretical Foundation and Hypotheses

The present study is based on fundament theoretical foundations—psychological principles that are basic to human behavior (unchanging). As such, I felt it was important to reference some of the relevant original seminal works.

3.1. Confirmation Bias

People have a need to see the world as consistent; this desire is theorized to motivate them toward confirmatory evidence. For example, Koriat, Lichtenstein, and Fischhoff (1980) have explained that people tend to seek examples that support their position or belief, to interpret information in ways that are biased towards those beliefs, and to give such information greater weight; and conversely, not to seek or even avoid information that is unsupportive of their positions, and to be less receptive to it. This leads to the *confirmation bias*: "unwittingly… seeking or interpreting evidence in ways that are partial to existing beliefs or expectations, [and thus confirming] a hypothesis in hand" (Nickerson, 1998, p. 175). When related to the biasing effect of previous performance information on subsequent evaluations, this phenomenon is also referred to as context effects, assimilation effects, or performance cue effects (Baltes and Parker, 2000b; Sumer and Knight, 1996). It has been found that a confirmatory bias can be so powerful that simply coming up with a specific hypothesis may be enough to increase one's confidence in it (Koehler, 1991). The central theme linking these similar effects is that an initial impression or perception of others may biasedly impact subsequent judgments and decision making. In line with these concepts, I propose that a person's reputation could create in others an initial expectation or belief that could trigger a confirmation bias and subsequent effects.

3.2. Model of Performance Appraisal

I also used DeNisi et al.'s (1984) model of performance appraisal to explain how a reputation might impact the three stages of the model (gathering and encoding, recall, and evaluating information), and thus bias SET performance ratings. In the first stage, holding to a reputation may cause raters to engage in a limited *search* or distorted *encoding*. Specifically, raters may spend less time observing the individual (DeNisi, et al., 1984), observe only selected portions of actual behavior that confirm initial positions (Jonas, et al., 2001), or biasedly interpret ambiguous information to further polarize their judgments instead of softening them (Nickerson, 1998). This is consistent with the proposition that high levels of status or reputation may cause stakeholders to interpret those persons' actions differently (Bednar, Love, and Kraatz, 2015). Specifically, scholars have suggested that leaders with positive reputations are afforded more trust, receive less monitoring, and are held to lower accountability standards (e.g., Hall et al., 2004). For example, researchers have found that organizations with a positive reputation were able to engage in deviant behavior without a reputational penalty (Deephouse and Carter, 2005), and in laboratory settings, fabricated information caused persons with a positive manipulation to be rated more favorably than were persons with a negative manipulation (e.g., Johnson, et al., 2002; Towler and Dipboye, 2006).

During the second stage (*recall*), some research indicates that when making judgments, people tend to remember more of others' behaviors that they expected (Nickerson, 1998). For example, in one study, participants recalled evidence that favored their theories as being more consistent than it actually was, and most were unable to recall the inconsistent evidence they had been provided (Kuhn, 1989). Thus, it is reasonable that students holding a positive reputation toward a professor may not recall as many poor teaching behaviors, and instead may recall average or good teaching as being

very good. In the third stage at the time of rating, a pre-held opinion may impact peoples' *evaluation* of information by biasing how they interpret and to what they attribute the behaviors they have observed (Miller and Turnbull, 1986). For example, research has found that identical average performance was judged as success and attributed to personal causes when enacted by persons for whom participants had a favorable view, but was discounted when enacted by "unfavorable persons" (Nickerson, 1998).

3.3. Belief Perseverance Effect

As people interact with others over time, this provides relevant performance information, so the accuracy of their judgments of others should increase. However, if individuals instead cling to initial categorizations, their perceptions may persist. This can result in the *belief perseverance effect*, an "individual's biased response to information in order to maintain an existing belief; a cognitive strategy of endorsing or soliciting confirming evidence but reacting more critically toward or even rejecting disconfirming information" (Jelalian and Miller, 1984, p.25). Scholars have recently tested this theory, proposing that organizations with a positive reputation may have a reservoir of goodwill that prompts stakeholders to be more lenient towards them and to give the organization the benefit of the doubt following a "wrong-doing" by the organization (Zavyalova, et al., 2016). For example, studies have found that following unexpected negative earnings reports, that high-reputation organizations experienced fewer short-term stock performance declines (Pfarrer et al., 2010); and following company downsizing, they had a smaller loss in their social approval compared to other firms (Love and Kraatz, 2009). Similarly, this effect could cause a reputation self-fulfilling prophecy (SFP) loop as certain employees gain a reputation as being ''on the fast track" influencing them to be promoted partly based on reputation, which gives them a more powerful reputation due to fast promotion, which leads to more promotions (Zinko, et al., 2007). Therefore, consistent with the theory and empirical evidence detailed above, and with an experimental design and setting to extend previous limitations, I hypothesize that:

H-1a: An initial *positive instructor reputation* will persist over time and will cause an *inflated performance rating*.

H-1b: An initial positive *course reputation* will persist over time and will cause an *inflated course rating*.

3.4. Halo Error

When individuals are categorized, it may be the label that is used to select and recall confirming information that further stabilizes previously formed impressions (Johnson, et al., 2002). Such labeling may cause a general impression halo, "whereby a rater's overall impression of a ratee leads the rater to evaluate all aspects related to performance consistent with this general evaluation" (Balzer and Sulsky, 1992, p.976). Thus, a *halo effect* occurs when a rater "fails to discriminate among conceptually distinct and independent aspects of a ratee's performance" (Pulakos, Schmitt, and Ostroff, 1986, p.29). One's reputation may invoke such labeling and cause a halo error which extends to ancillary areas connected with the ratee. Marsh (1984) indicated that teaching evaluations could be biased when a rating in one category affected the ratings of other independent areas. The present experiment tests this concept within the framework of a positive reputation. Therefore, I hypothesize:

H-2: A *positive instructor reputation* will cause ratings of ancillary areas (*course*, *grading*, and *materials*) to be biased upward due to a positive halo error.

3.5. Amount Learned: Actual Versus Perceived

Consistent with the discussion so far, research has also shown that other things can impact students' perceptions of learning other than the amount they actually learned (Downinga, et al., 2018). For example, in one study the increased cognitive effort associated with active learning methods biased students' estimations of the amount they learned. Specifically, students in active learning classrooms learned more than those in passive instruction classes, but their perceptions of learning were lower (Deslauries, et al., 2019). It has also been found that students' perceptions of learning can be influenced by (correlated with) their perceptions and evaluations of their instructor (Deslauries, et al., 2019). This supports the premise that if something causes students' evaluations of their instructor to be inflated, then their evaluations of the amount they learned may also be inflated. Similarly, the halo effect would support that students' holding a positive instructor reputation may overshadow and thus bias their related but distinct estimation of how much they learned in the instructor's course. These perceptual phenomena highlight the disconnect between what one believes and reality. In this case, holding instructor biases would not be expected to impact the amount students actually learn. Therefore, based on the theoretical and empirical support provided, I hypothesize that:

H-3: A *positive instructor reputation* will cause upwardly biased evaluations *(perceptions) of the amount learned* in the course.

H-4a: A *positive instructor reputation* will not be related to the *actual amount learned* in the course.

That being said, the SFP known as the "Messiah Effect," supports the opposite impact of instructor reputation on the amount learned. The *messiah effect* occurs when followers' expectations of their leader have an influence upon the followers' behavior or achievement level (Eden, 1990). In essence, people may act consistent with the views they have of their superior, thus causing their expectations to be fulfilled. In the present case, if a positive reputation leads students to believe that their teacher is wonderful (can help them learn), then this theory predicts they will pay closer attention in class, take good notes, participate in discussions, be diligent with assignments and projects, prepare well for exams, and otherwise act in ways that cause them to learn more and thus to fulfill their prophecy. Therefore, I also offer the competing hypothesis of the effect on actual learning:

H-4b: A *positive instructor reputation* will be related to the *actual amount learned* in the course.

Consistent with the halo error discussion above, when rating how much they learned in a course, students may focus on something besides their degree of learning. The extent to which a positive reputation is present may bias their accurately assessing the amount they've learned. Essentially the student could default to the logic that if I had a great professor this semester, I must have learned a lot. Thus, if students' assessment of learning is biased, it will not correlate with the amount they actually learned. Some authors have argued that novices in particular are poor at judging their actual learning, and thus may rely on inaccurate metacognitive cues such as their evaluation of the instructor when attempting to assess their own learning (Deslauries, et al., 2019).

This line of reasoning is also supported by an extrapolation of the Transitive Property of mathematics (if A = B, but A ≠ C, then B ≠ C), where A = positive reputation, B = perception of learning, and C = actual learning. Specifically, if a positive reputation upwardly biases perceptions of the amount learned, but holding a positive reputation does not help one learn more, then perceptions of the amount learned will not be related to the actual amount learned. For example, in one study, the instruction method was positively related to actual learning (active learned more), but the instruction method was not significantly related to the student's perceptions of how much they had learned (Deslauries, et al., 2019). By extension, then, the actual amount students learned should not have been related to their perceptions of learning. Thus, I hypothesize that:

H-5: When a reputation bias is present, students' *perception of the amount learned* in the course will not be related to the *actual amount learned* in the course.

4. Method

4.1 Sample, Design, and Measures

The participants were 60 1st-year Masters of Business Administration (MBA) students who were assigned to one of two sections of the same required course (with two high-performing professors). The MBA office randomly assigns MBA students to sections of courses after stratifying them according to gender, years of work experience, and foreign nationality. These MBA students were college graduates, 54% were male, and they had a mean age of 27.4 and 4 years of work experience. Seventy-five percent were from the U.S. and 25% from other countries. The field study used a naturally occurring experimental situation over a four-month period with several data collection points. The intervention was enacted the week before courses began, and the reputation intervention manipulation check and demographic information were gathered on a survey two days before the beginning of courses. Throughout their MBA program students periodically had the option to complete surveys for various research projects. The learning measures were completed at two months and four months as a function of the course, and all student's ratings were done after the semester (four months).

The *Instructor Reputation* and *Course Reputation* were measured on the survey. Participants were asked to recall what they had heard about each professor's reputation, experience, and qualifications and from whom they had heard it. Then, using a five-item, five-point scale where 1=*very unfavorable*, 3=*neutral or no reputation*, and 5=*very favorable*, participants indicated how favorably what they had heard reflected upon each professor's reputation, experience, qualifications, credentials, and expertise (α=.94). Participants were similarly asked to recall what they had heard about the course, and using the same scale, to rate how favorable that information was about the course.

The student evaluation variables were measured using the University's SET instrument with a 5-point response scale ranging from 1=*F* to 5=*A*. The *Teacher evaluation* assessed the professors' performance with six items such as "Instructor's daily preparation for class" (α=.89). The *Materials evaluation* consisted of two items which assessed the "Readability" and "Quality" "of textbooks and other instructional materials" (α=.89). The *Grading evaluation* was comprised of three items such as rating the "Clarity of examinations" (α=.89). The *Course evaluation* was measured with three items such as "Clarity of course objectives" (α=.85). Students *Perceived learning* was measured by rating the "Amount learned in this course." Finally, *Actual learning* was measured using the total points earned in the course (composite of two examinations, course project, cases, and participation). Scores from such course components are generally accepted measures of learning performance (Armstrong, 1998). Both sections used the same project, cases, exams, and grading criteria. As well, both professors graded a subset of exams and assignments from the other's section to ensure a comparable level of grading (average inter-rater agreement r_{wg}=.93).

4.2. Naturally-Occurring Intervention

Third-party signaling can influence peoples' perceptions through both formal channels (e.g., awards and recognitions), and informal word-of-mouth information (Dineen and Allen, 2016). Word-of-mouth is defined as an independent, interpersonal communication about a given target (Bone, 1995). It is a social phenomenon that occurs between people in an informal manner with a target-independent information source—i.e., a source not under the direct control of the target, nor any self-interest in promoting the target (Wirtz and Chew, 2002). The theory posits that the impact of word-of-mouth in creating a reputation effect on people's perceptions and behaviors is a function of credibility, relationship, comparability, and timing (Lau and Ng, 2001; Van Hoye and Lievens, 2009). First, credibility is influenced by the independent nature of the source (compared to the questionable veracity of self-serving comments), and the perceived level of expertise of the source—knowledge and experience with the target (Bansal and Voyer, 2000). The second consideration is the closeness, or "tie strength" of the social relationship between the recipient and the source of word-of-mouth information (Brown and Konrad, 2001). Next, information on the relative standing of a target compared with other targets is thought to create a stronger or more permanent reputation (Bangherter, Roulin, and Konig, 2012). Finally, word-of-mouth reputations may have greater influence the earlier they occur, due to recipients' lack of information regarding the target (Zinko, et al., 2007). For example, researchers found that receiving positive employment information through word-of-mouth early in the recruitment process was associated with organizational attractiveness and actual application decisions (Van Hoye and Lievens, 2009).

MBA students in this program have several pre-fall formal and informal social gatherings that occur during the week before classes begin. During these events the new 2nd-year students assume the role of orienting and socializing the new 1st-year students including informing them about their professors and courses. It has been suggested that when individuals share their impressions of individuals with others (such as during these MBA gatherings), that these impressions may influence others' attitudes, expectations, and behaviors (Bromley, 1993). Thus, it was believed that these upcoming 2nd-year students might create a naturally occurring reputation intervention by passing on professor and course information to the new 1st-year MBA students. All of the conditions specified in the theory above were met to create the naturally-occurring intervention in the present study. These gatherings would be (1) early-on in the program, (2) the 2nd-year MBAs would be a very credible source (independent and expert) with a favorable relationship with first years created by their sharing an MBA experience at the same school, and (3) they would provide information with relative comparisons among professors. In fact, Smither et al., (1988) found that students rated *information from previous semester's students* as the most credible source of professors' teaching skills—higher than information from an academic advisor, alumni, other professors, or even members of their fraternity/sorority.

Previous to the commencement of fall semester, I designed the present experiment to test the extent to which a naturally occurring positive reputation might create a confirmation bias and belief perseverance effects that subsequently influenced evaluations of performance, the course, and learning. Therefore, I selected a 1st-year MBA course with sections taught by two professors whom I knew were generally well-accepted, and could be expected to have positive things said about them by the 2nd-year MBAs.

4.3. Combining Sections for Analysis

In order to analyze the data from the two sections together, I reviewed the literature to discover any factors that might impact SETs to assure that these were sufficiently equivalent between the two sections. First, everything about the course was sufficiently identical between the two sections. It was the same course (strategy) and course level

(Constanda, Clarke, and Morgan, 2018), and similar in class size—31 versus 29 (Crittenden, Norr, and LeBailley, 1975). Since the two professors had jointly developed the course, the sections used the identical syllabus, textbooks, and articles (Abrami and d'Apollonia, 1990), had the same PowerPoint slides, lectures, and in-class activities (Sheehan and Duprey, 1999), and included the same workloads: assignments, cases, exams, and semester project (Abrami and d'Appolonia, 1990). It was also a required course (Pan, et al., 2021) with sections that met at the same time, in the same building, and in identical classrooms (Marsh, 1984). In addition, I used a t-test to compare the actual amount learned (total course points for exams, project, etc.) and found no significant difference between the sections.

Second, there were no significant differences in student characteristics between the sections, including gender (Tatro, 1995), age, ethnicity, undergraduate major (Bachen, et al., 1999), and number of years of work experience. Random assignment to the sections likely contributed to this, and should also have created equivalence between the sections in students' previous exposure to or interest in the subject matter (Marsh, 1984; Phillips, 1999). Third, professor characteristics that might impact student ratings were sufficiently equivalent. Although the professors were not the same gender (Constanda, Clarke, and Morgan, 2018), they were the same ethnicity (Phillips, 1999), similar in age—39 and 41 (Marsh, et al., 1997), both were professors of strategy with experience teaching at the undergraduate and graduate levels, and with similar years of experience teaching (8 and 10). Finally, to examine any impact from all other factors that might create a difference in performance evaluations between the professors, such as their teaching ability, personality, teaching style, looks, etc., with their consent I accessed and compared the teaching evaluations of the two professors. First, I found no significant difference between their SETs from the previous three years (means of 4.63 and 4.65). Second, I performed t-tests to compare the end-of-semester student evaluations used in the present study (professor, course, perception of amount learned, etc.), and found no significant difference. Thus, given all of the course, student, and professor equivalences between the sections, it is reasonable to combine the data to test the hypotheses.

5. Results

Descriptive statistics, reliabilities, and intercorrelations are provided in Table 1. T-tests were used for the manipulation checks, and intercorrelations were used to test the hypotheses. First, the manipulation checks tested the extent to which a naturally occurring positive reputation had been created for the students' professor and course (whether professor's reputation and course reputation were significantly greater than neutral—i.e., 3 on the scale). Results indicated that both the professor reputation ($t = 8.66$, $p < .000$) and course reputation ($t = 6.31$, $p < .000$) were significantly favorable. This validated that a naturally occurring positive reputation manipulation had occurred.

To test the hypotheses, the first analysis examined whether the positive reputation persisted over time to create a bias on evaluations of professor performance. This was verified as there was a significant correlation between the level of initial positive instructor reputation held and evaluations of instructor performance four months later ($r = .64$, $p < .001$). Thus, H-1a was supported. A similar analysis revealed that the initial positive course reputation had no effect on the subsequent rating of the course ($r = .19$, *ns*), so H-1b was not supported.

Next, I tested whether the halo effect was present among the ratings of ancillary areas, and of any impact on these ratings created by the positive instructor reputation. There are a variety of methods used in the literature to calculate and assess the presence of halo (Balzer and Sulsky, 1992). Consistent with the paper's conceptualization of halo I used the most common method—intercorrelations between variables (Pulakos, et al., 1986). First, I calculated the intercorrelation between the ratings of professor's performance and the ratings of the ancillary areas—course, grading, and materials. Student evaluations of each of these areas were found to be highly correlated with the evaluations of instructor performance: course evaluation ($r = .72$, $p < .001$), grading evaluation ($r = .74$, $p < .001$), and materials evaluation ($r = .68$, $p < .001$). Second, to examine any biasing effect from the positive reputation, I calculated the correlations between the level of positive reputation held and subsequent ancillary evaluations. The reputation significantly impacted each of these: course evaluation ($r = .45$, $p < .001$), grading evaluation ($r = .46$, $p < .001$), and materials evaluation ($r = .40$, $p = .002$). These analyses revealed that students engaged in halo errors, and that although they in fact experienced the identical course, grading procedures, and course materials, they biasedly rated them higher when they held a more positive reputation of their instructor. Thus, H-2 was supported.

I then examined whether holding a positive reputation of one's instructor might upwardly bias students' subsequent estimations of the amount they had learned in the course. Analysis confirmed this hypothesis (H-3)—$r = .43$, $p < .001$. Inflated views of their professor worsened the accuracy of the students' judgment and caused them to believe they had learned more. I next tested the competing hypothesized relationships between a positive instructor reputation and the

actual amount learned in the course. Correlational results indicated there was no significant relationship (r = -.08, *ns*). In other words, having an initial highly favorable perception of their professor did not help students learn more (H-4a supported; H-4b not supported). Finally, I examined the relationship between students' estimation of the amount they learned and the actual amount they learned. Since I had hypothesized that when a reputation bias is present that student's estimations of the amount they learned would be inflated, I then hypothesized that there would not be a relationship between these perceptions and the actual amount they had learned. Analysis of the data confirmed this (H-5)—(r = -.02, *ns*).

6. Discussion and Conclusions

Performance evaluations have a vital impact on people's employment and career success, and influence organizational effectiveness. As such, I examined a naturally occurring field experiment within academia to test any rating bias and belief perseverance caused by a positive reputation. This longitudinal field testing provided important, next-step contributions to support and expand our understanding of several theories and models, and to build on previous laboratory studies that used short-term, artificially generated targets and biases. Through the combined lenses of the performance cue effect, the cognitive model of performance appraisal (e.g., DeNisi, et al., 1984), and the personal reputation literatures (e.g., Ferris, et al., 2003), the results indicated that a person's reputation can serve as a preconceived notion that disrupts subsequent accurate rating processes. Specifically, in a real work setting (academia), initial information passed to raters formed a positive reputation that persisted for 16 weeks in the midst of actual behaviors they experienced, and biasedly inflated their performance evaluations and ancillary ratings.

These findings are not a confound of the actual level of instructor performance. In other words, the large correlation between reputation and performance ratings is not because these were great professors whose high SET accurately reflect their high performance. No, it is vital to remember that all students experienced the same instructor performance. So, if no bias were operating, all evaluations would be similar regardless of the level of the students' initially held reputation for the professor. But that was not the case; the reputation drastically inflated students' ratings of performance—an effect of over one and a half standard deviations (d = 1.67). This should raise some concern as it demonstrates a legitimate contaminant of the validity of performance ratings (SET) when a reputation is present.

Next, the results also contribute to the literature by demonstrating that a halo error could occur due to a positive ratee's reputation. Ratings of aspects of the course that were only tangentially related to the instructor were biased upward. For example, the evaluation of the course was not affected by the reputation of the course as might be expected, but instead was affected by the reputation of the instructor. Even more empirically blatant were individuals' biased evaluations of the materials and the grading in the course. All students experienced the same course grading structure and grading methods, and literally had the identical textbook and instructional materials (videos, cases, articles, etc.). Thus, it would be illogical if students were to see and think about these resources differently depending upon a reputation they held of the instructor. But they did. They saw and evaluated these identical things almost one standard deviation differently (d = .97) to the extent they held a positive reputation of the instructor. This is powerful evidence of a Halo Effect. The students did not effectively distinguish among the things they were rating. Their evaluations of the course, grading, and materials were partly a measure of their positively biased opinion of the instructor.

Inasmuch as student learning is a primary outcome of course instruction and the teaching process, the findings in this area are particularly important. The results highlighted that when students held a positive reputation of the instructor, it caused them to think they had learned more in the course; but it did not cause them to learn any more. This is because the effect of the positive reputation on the students was perceptual not actual. In this case, over-estimating the quality of their teacher provided no benefit; it only created a confound that lessoned their ability to perceive and judge accurately. At first brush this seems to contradict the Messiah Effect hypothesized in H-4b (that higher expectations of your leader will result in greater follower performance). However, the theory specifies that in order for the Messiah Effect to operate, the follower (student) needs to be consciously and actively thinking about and holding that positive view/expectation of their leader (teacher) throughout interactions, in a way that causes the follower to do things differently (better, more)—Eden, 1990. In this academic course, the positive reputation perception was likely not actively salient in the students' minds, or at least not in a way to cause them to work harder and more conscientiously throughout the course. They may not have consistently thought about the professor's reputation until called upon to make an evaluative judgement at the end of the semester. Or perhaps they thought that since the teacher was so great, that they would learn without having to work harder to do well in the course.

Relatedly, there was no relationship between how much students thought they had learned, and the amount they had actually learned. This calls into question the usefulness and validity of students' self-assessment of learning. The results showed that when a reputation effect is present, student's evaluation of their learning is not a true measure of their estimate of how much they learned; rather, it is partly a reflection of their pre-semester belief about their instructor's quality. In other words, the perceived learning measure is capturing the positive reputation bias rather than being an accurate perception of their amount learned. This is verified through examining the extremely high correlation between Perception of Amount Learned and the Evaluation of the Instructor (r = .70, p =.000)—an effect of almost two standard deviations (d = 1.96).

Taken together, the present results provide additional evidence of the potential dubious accuracy and usefulness of SETs in evaluating faculty teaching performance, and highlight the importance of universities and other organizations pro-actively monitoring and managing situations where a reputation may be biasing judgements. Given these results that show how susceptible SETs are to bias, administrators should exercise caution when using and interpreting SETs, and the extent to which they rely upon them in making decisions (Constanda, Clarke, and Morgan, 2018). Researches have recommended that SETs should be tailored to reflect the idiosyncrasies of different teaching pedagogies and academic disciplines, and actively involve students and faculty in their development (Pan, et al., 2021). In addition, data for important decisions such as faculty annual reviews, course assignments, tenure, and promotion, should be augmented with other sources and methods. These might include peer faculty in-class observations, examination of course syllabi and course materials and assignments used, as well as evaluating other facets of teacher performance such as care and empathy in meeting students' needs, out-of-class time spent helping students, and so forth.

Next, human resource experts have indicated that the corporate world shares similar HRM challenges with academic institutions, and since principles applied in academic settings are relevant in the business world, much can be learned within and from academic institutions (Lorange, 2006; Ulrich, 2006). Relatedly, applied HRM work is germane to the broader area of developing and managing the human capital element of organizations (e.g., Hesketh, 2014). Thus, future research could examine implications for professional and institutional practice. For example, what might be the extent to which the reputation biasing effects found in the present study generalize to employee ratings within business contexts? Two general differences between standard performance appraisals in non-academic industries and SET are that most performance appraisals (1) are done by those in a higher position (supervisor), and (2) occur repeatedly over time by the same rater. On the other hand, SET are completed by subordinates (students) engaging in a one-time evaluation for a given professor. That being said, SET still serve as *the* direct appraisal of professors' teaching performance—any subsequent upper-level evaluations are primarily based on SET ratings (Waldman, 2008). As well, there are many one-time evaluations that occur within non-academic organizations that may be impacted by persons' reputations. For example, when employees vote on given project proposals, strategy directions, or similar decisions, they may be influenced by the reputation of proponents and/or opponents of the given proposals. Thus, additional research should examine reputational effects on appraisals in other work settings, and the impact on given project decisions due to the reputation of persons associated with the projects. Future research could also target processes of the cognitive model of performance appraisal to further clarify at what points the biasing effect of a reputation takes place. In addition, future research should examine ways universities and other organizations can prevent or minimize reputations biasing effects on performance ratings. Initial laboratory studies have found that participants may correctly take into account disconfirming information when they are taught the belief perseverance phenomenon (Misra, 1992; Jenning, et al., 1981), engage in halo bias training (Baltes and Parker, 2000b), or do structured recall exercises before rating (Baltes and Parker, 2000b). Research should investigate the extent to which such practices generalize to field settings.

Experts have stressed that the ultimate goal of performance appraisal research should be to improve rating accuracy (e.g., DeNisi, et al., 1984). The present experiment aids this cause by documenting the presence of and alerting organizations to the potential biasing effect of reputations on performance ratings. The results are particularly germane to academia where even after four months of experiencing the same course, a naturally-occurring positive reputation caused students' ratings of professor, grading, course materials, and learning to be significantly inaccurate. Thus, this study provides additional empirical evidence to the growing literature that urges caution with the extensive use of student evaluations of teaching as a measure and evaluation tool for assessing instructor's performance.

References

Abrami P and d'Apollonia S (1990). The dimensionality of ratings and their use in personnel decisions. *New Directions for Teaching and Learning*, 43: 97-111.

Armstrong JS (1998). Are student ratings of instruction useful? *American Psychologist*, 53: 1223-1232.

Bachen C.M, McLoughlin MM, and Garcia SS (1999). Assessing the role of gender in college students' evaluations of faculty. *Communication Education*, 48: 193-210.

Baltes BB and Parker CP (2000a) Understanding and removing the effects of performance cues on behavioral ratings. *Journal of Business and Psychology*, 15: 229-246.

Baltes BB and Parker CP (2000b) Reducing the effects of performance expectations on behavioral ratings. *Organizational Behavior & Human Decision Processes*, 82: 237-267.

Balzer WK and Sulsky LM (1992). Halo and performance appraisal research: A critical examination. *Journal of Applied Psychology*, 77: 975-985.

Bangherter A, Roulin N, and Konig CJ (2012) Personnel selection as a signaling game. *The Journal of Applied Psychology*, 97(4): 719-738.

Bansal HS and Voyer PA (2000) Word-of-mouth processes within a services purchase decision context. *Journal of Service Research*, 3(2): 166-177.

Boyd BK, Bergh DD, and Ketchen DJ Jr (2010) Reconsidering the reputation-performance relationship: A resource-based view. *Journal of Management*, 36(3): 588-609.

Bromley DB (1993) Reputation, image and impression management. New York: Wiley.

Brown DW and Konrad AM (2001) Granovetter was right: The importance of weak ties to a contemporary job search. *Group and Organization Management*, 26(4): 434-462.

Chow IH (2004) Human resource management in China's township and village enterprises: Change and development during the economic reform era. *Asia Pacific Journal of Human Resources*, 42: 318-335.

Constanda RL, Clarke N, and Morgan M (2018) An analysis of the relationships between management faculty teaching ratings and characteristics of the classes they teach. *The International Journal of Management Education*, 16(2): 166-179.

Crittenden KS, Norr JL, and LeBailley RK (1975) Size of university classes and student evaluations of teaching. *Journal of Higher Education*, 46: 461-470.

Deephouse DL and Carter SM (2005) An examination of differences between organizational legitimacy and organizational reputation. *Journal of Management Studies*, 42(2): 329-360.

DeNisi AS, Cafferty TP, and Meglino BM (1984) A cognitive view of the performance appraisal process: A model and research propositions. *Organizational Behavior and Human Performance*, 33: 360-396.

Deslauries L, McCarty LS, Miller K, et al. (2019) Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences*, 116: 19251-19257.

Dineen BR and Allen DG (2016) Third party employment branding: Human Capital inflows and outflows following "best places to work" certifications. *Academy of Management Journal*, 59(1): 90-112.

Downing JA, Aiken D, McCoy D, et al. (2018) Collaborative course development: A comparison of business and non-business students' perceptions of class experience. *The International Journal of Management Education*, 16(2): 256-265.

Eden D (1990) *Pygmalion in management: Productivity as a self-fulfilling prophecy*. Lexington, MA: Lexington Books.

Ferris GR, Blass FR, Douglas C, et al. (2003). Personal reputation in organizations. In J. Greenberg (Ed.) Organizational behavior: The state of the science. Mahwah, NJ: Erlbaum.

Fuller MA and Delorey R (2016) Making the choice: University and program selection factors for undergraduate management education in Maritime Canada. *The International Journal of Management Education*, 14(2): 176-186.

George G, Dahlander L, Graffin SD, et al. (2016) Reputation and status: Expanding the role of social evaluations in management research. *Academy of Management Journal*, 59(1): 1-13.

Hesketh A (2014) Managing the value of your talent: A new methodology for human capital measurement. Chartered Institute for Personnel Development Research Report, 1-93.

Hubbard BA and Smith C (2003) The growth of full-time nontenure-track faculty: Challenges for the union. *American Federation of Teachers Report*, Item number 36-0700.

Jelalian E and Miller AG (1984) The perseverance of beliefs: Conceptual perspectives and research developments. *Journal of Social and Clinical Psychology*, 2: 25-56.

Jenning DL, Lepper MR, and Ross L (1981) Persistence of impressions of personal persuasiveness: Perseverance of erroneous self-assessments outside the debriefing paradigm. *Personality and Social Psychology Bulletin*, 7: 257-263.

Johnson DE, Erez A, Kiker DS et al. (2002) Liking and attributions of motives as mediators of the relationships between individuals' reputations, helpful behaviors, and raters' reward decisions. *Journal of Applied Psychology*, 87: 808-815.

Jonas E, Schulz-Hardt S, Frey D, et al. (2001) Confirmation bias in sequential information search after preliminary decisions: An expansion of dissonance theoretical research on selective exposure to information, *Journal of Personality and Social Psychology*, 80: 557-571.

Kilduff M, Crossland, Tsai W, et al. (2016) Magnification and correction of the acolyte effect: Initial benefits and ex post settling up in NFL coaching careers. *Academy of Management Journal*, 59(1): 352-375.

Koehler DJ (1991) Explanation, imagination and confidence in judgment. *Psychological Bulletin*, 110: 499-519.

Koriat A, Lichtenstein S, and Fischhoff B (1980) Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, 6:107-118.

Kuhn D (1989) Children and adults as intuitive scientists. *Psychological Review*, 96: 674-689.

Lau GT and Ng S (2001) Individual and situational factors influencing negative word-of-mouth behavior. *Canadian Journal of Administrative Sciences*, 18(3): 163-178.

Lorange P (2006) A performance-based, minimalist human resource management approach in business schools. *Human Resource Management*, 45: 649-658.

Love EG and Kraatz M (2009) Character, conformity, or the bottom line? How and why downsizing affected corporate reputation. *Academy of Management Journal*, 52(2): 314-335.

Marsh HW (1984) Students' evaluations of university teaching: Dimensionality, reliability, validity potential biases, and utility. *Journal of Educational Psychology*, 76: 707-754.

Marsh HW, Hau K, Chung C, et al. (1997) Students' evaluations of university teaching: Chinese version of the Students' Evaluations of Educational Quality Instrument. *Journal of Educational Psychology*, 89: 568-572.

Martell RF and Willis CE (1993) Effects of observers' performance expectations on behavior ratings of work groups: Memory or response bias. *Organizational Behavior and Human Decision Processes*, 56: 91-109.

Miller DT and Turnbull W (1986) Expectancies and interpersonal processes. *Annual Review Psychology*, 37: 233-256.

Misra S (1992) Is conventional debriefing adequate? An ethical issue in consumer research. *Journal of the Academy of Marketing Science*, 20: 269-273.

Munz DC and Munz HE (1997) Student mood and teaching evaluations. *Journal of Social Behavior & Personality*, 12: 233-242.

Nickerson RS (1998) Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2: 175-220.

Pan G, Shankararaman V, Koh K, et al. (2021) Students' evaluation of teaching in the project-based learning programme: An instrument and a development process. *The International Journal of Management Education*, 19(2): 100501.

Pfarrer MD, Pollock TG, and Rindova VP (2010) A tale of two assets: The effects of firm reputation and celebrity on earnings surprises and investors' reactions. *Academy of Management Journal*, 53(5): 1131-1152.

Phillips SB (1999) *Student evaluation of faculty instruction: Inflated results and student feedback*. Dissertation Abstracts International, 60, 1478.

Pulakos ED, Schmitt N, and Ostroff C (1986) A warning about the use of a standard deviation across dimensions within ratees to measure halo. *Journal of Applied Psychology*, 71: 29-32.

Raub W and Weesie J (1990) Reputation and efficiency in social interactions: An example of network effects. *American Journal of Sociology*, 96: 626-654.

Rindova VP, Williamson IO, Petkova AP, et al. (2005) Being good or being known: An empirical examination of the dimensions, antecedents, and consequences of organizational reputation. *Academy of Management Journal*, 48(6):1033-1049.

Sheehan EP and DuPrey T (1999) Student evaluations of university teaching. *Journal of Instructional Psychology*, 26: 188-193.

Smither JW, Reilly RR, and Buda R (1988) Effect of prior performance information on ratings of present performance: Contrast versus assimilation revisited. *Journal of Applied Psychology*, 73: 487-496.

Sumer HC and Knight PA (1996) Assimilation and contrast effects in performance ratings: Effects of rating the previous performance on rating subsequent performance. *Journal of Applied Psychology*, 81: 436-442.

Tatro CN (1995) Gender effects on student evaluations of faculty. *Journal of Research & Development in Education*, 28: 169-173.

Towler A and Dipboye R (2006) Effects of trainer reputation and trainees' need for cognition on training outcomes. *The Journal of Psychology*, 140, 549-564.

Turban DB and Cable DM (2003) Firm reputation and applicant pool characteristics. *Journal of Organizational Behavior*, 24(6): 733-751.

Turban DB, Forret ML, and Hendrickson CL (1998) Applicant attraction to firms: Influences of organization reputation, job and organizational attributes, and recruiter behaviors. *Journal of Vocational Behavior*, 52: 24-44.

Ulrich D (2006) Academic application is not an oxymoron. *Human Resource Management*, 45: 663-665.

Van Hoye G and Lievens F (2009) Tapping the grapevine: A closer look at word-of-mouth as a recruitment source. *The Journal of Applied Psychology*, 94(2): 341-352.

Waldman DA (2008) Readdressing the age-old question: What to Study? *Academy of Management Learning and Education*, 7: 153-157.

Zavyalova A, Pfarrer MD, Reger RK, et al. (2016) Reputation as a benefit and a burden? How stakeholders' organizational identification affects the role of reputation following a negative event. *Academy of Management Journal*, 59(1): 253-276.

Zinko R., Ferris GR, Blass FR, et al (2007). Toward a theory of reputation in organizations. *Research in Personnel and Human Resources Management*, 26: 163–204.

Table 1

Means, standard deviations (SD), scale reliabilities, and intercorrelations among study variables

Variables	Mean	SD	1	2	3	4	5	6	7	8
1. Teacher reputation	3.82	(.73)	**.94**							
2. Course reputation	3.74	(.88)	.47	--						
3. Teacher evaluation	4.46	(.63)	.64	.33	**.89**					
4. Materials evaluation	3.73	(1.07)	.40	.24	.68	**.89**				
5. Grading evaluation	3.97	(.89)	.46	.21	.74	.66	**.89**			
6. Course evaluation	4.08	(.81)	.45	.18	.72	.52	.67	**.85**		
7. Perceived learning	4.19	(.92)	.43	.27	.70	.51	.65	.76	--	
8. Actual learning	90.82	(3.00)	-.08	-.20	-.08	.12	-.07	-.04	-.02	--

n = 59 - 60. Coefficient alpha reliability estimates are on the diagonal.

$r \geq .27$, $p < .05$ $r \geq .43$, $p < .00$