Journal of Applied Research in Memory and Cognition 9 (2020) 170–174

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

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

Commentary

If Teaching Evaluations Don't Measure Learning, What Do They Do?

Daniel M. Oppenheimer*
Carnegie Mellon University, United States

Mary B. Hargis
Texas Christian University, United States

Because student evaluations of teaching are so commonly used to assess instructor performance, there has been a great deal of recent attention on the extent to which such evaluations are reliable and valid. Carpenter, Tauber, and Witherby (2020) have synthesized this literature to compellingly demonstrate that teaching evaluations are not predictive of student learning. This work is highly relevant to learners and instructors in several ways, not the least of which is that decisions about instructors' compensation and career advancement often depend upon these evaluations. Reliance on such evaluations creates fundamental unfairness in the promotion and tenure process, as factors such as age, ethnicity, and gender influence evaluations, even when the teaching itself is held constant. Moreover, some characteristics that students prefer are actually negatively correlated to their longer-term learning (Carpenter et al., 2020). This mismatch incentivizes teachers to engage in strategies that are known to be less effective for student learning, but which might nonetheless raise teaching evaluations.

In some sense, the flaws of teaching evaluations (by students) are inevitable given the cognitive mechanisms that underlie judgments of this nature. As Carpenter et al. (2020) argue, metacognitive errors are a key contributor to the lack of relationship between students' evaluations of teaching and later retention of the material. When students attempt to evaluate the class, they likely do not have direct metacognitive access to what they have learned. It has been well documented that when people learn new information, they believe that they actually knew it all along (Fischhoff, 1975)[1] . Moreover, unless students directly interrogate their memories for specific information, they are unlikely to be aware of what they have forgotten (Hargis & Oppenheimer, under review). In fact, learners are unlikely to even think to interrogate their memory for information that they do not know. Asking people to evaluate how much they have learned is an incredibly difficult question.

From a decision theoretic perspective, when people are faced with such difficult questions, they often engage in attribute substitution, exchanging the difficult judgment for an easier one (Kahneman & Frederick, 2005). In this case, given the difficulty of evaluating learning, the easier question might be something like "How much did you like this class?"[2] It is well known that judgments of liking are strongly influenced by fluency—that is, the subjective feeling of ease of processing (for a review, see Alter & Oppenheimer, 2009). This finding may explain why courses that feel easy, either through grade inflation or through avoiding desirable difficulties (which improve learning but reduce fluency; Bjork & Bjork, 2011), are more positively evaluated, independent of learning outcomes. This kind of holistic, hedonic judgment is also influenced heavily by stereotypes (c.f. Cuddy, Fiske, & Glick, 2007) and subject to a large array of biases (for a review, see Frederick & Loewenstein, 1999).

Given the above analysis, "fixing" teaching evaluations may be an insurmountable challenge. However, we would argue that while teaching evaluations may not correlate with student learning in a course, they may still have other positive benefits that are worth considering when developing teaching assessment policies.

[1] A similar bias is the curse of knowledge (Camerer, Loewenstein, & Weber, 1989) in which people who know a fact have trouble getting into the mindset of those who don't. The existence of this bias suggests that teaching evaluations from faculty peers may suffer from similar problems as teaching evaluations from students. In both cases, it is very difficult to evaluate the extent of knowledge that has been gained.

Author Note

Daniel M. Oppenheimer (Departments of Social and Decision Sciences and Psychology, Carnegie Mellon University). Mary B. Hargis, Department of Psychology, Texas Christian University.

Thanks to Nicole Oppenheimer for helpful feedback.

* Correspondence concerning this article should be addressed to Daniel M. Oppenheimer, Departments of Social and Decision Sciences and Psychology, Carnegie Mellon University, United States. Contact: oppenheimer@cmu.edu.

First, there may be some content within the qualitative (e.g., short-answer) components of students' evaluations that may be helpful for teachers. Of course, as Carpenter et al. (2020) document, teaching evaluations can be biased against women (Basow & Martin, 2012), instructors of color (Bavisi, Madera, & Hebl, 2010), and people with non-native accents (Sanchez & Khan, 2016). We do not suggest that comments about appearance, accent, or other biased content should be addressed by the instructor. However, a pattern of comments suggesting that the instructor speaks too quickly or that the font on their slides is too small to read could be addressed in a future semester and may benefit students. We agree with Carpenter et al. (2020) that qualitative evaluations may be potentially helpful for the instructors, insofar as students comment on aspects of the instruction that could truly improve student learning outcomes.

Second, factors that may not directly affect learning of course material may still prove valuable to students in other ways. For example, instructor enthusiasm may not help students retain knowledge (Carpenter et al., 2020). However, such an instructor could spark students' curiosity in the subject material or in learning more generally.

There is precedent for this pattern of results. Research on early childhood educational interventions has often observed "fade out" effects, in which demonstrated IQ and achievement gains from excellent teaching or small class sizes diminish over time and are no longer present in subsequent classes (for a review and meta-analysis of fade out effects, see Protzko, 2015). Indeed, in one classic study (the STAR project) of nearly 12,000 students who were randomly assigned to kindergarten classrooms, the quality of that classroom (as measured by class size and immediate student performance on tests) had a decreasing effect on student performance in the years following the class, until by eighth grade it had no measurable effect on test scores whatsoever. However, a subsequent longitudinal analysis of participants from the STAR program showed that the quality of one's kindergarten class affected students' likelihood of taking college entrance exams (Krueger & Whitmore, 2001), as well as the likelihood of attending college and their eventual earnings as adults (Chetty et al., 2011). These studies have shown that these positive outcomes are likely due to improvements to the students' non-cognitive skills, such as effort and initiative (Chetty et al., 2011). In other words, while students in high quality kindergarten courses didn't show persistent improvement in subsequent courses, students who had been inspired to work hard and be proactive nonetheless had better life outcomes.

Moreover, the extent to which students enjoy a course, even if that enjoyment is not correlated with learning, could encourage them to continue on to advanced coursework in that domain. For example, one study examined cadets at a military academy where they were randomly assigned to course schedules for general education coursework (Haggag, Patterson, Hope, & Feudo, under review). Some cadets were assigned to a 7:30am section for chemistry, English, math, or physics, while other sections were later in the day. While assignment to a 7:30 am class did not affect their learning or performance in that class, it did lead students to be 10% less likely to major in the subject, ostensibly because they found the class less enjoyable, and misattributed their lack of enthusiasm to their passion for the subject rather than to their fatigue. In other words, if a student enjoys a course (or doesn't enjoy a course), that experience may have downstream effects on outcomes that are not just measured by tests in a future course but by persistence in becoming an expert in that domain. Thus, while course evaluations may be driven more by hedonic impressions than actual learning, this is not necessarily a completely irrelevant consideration.

Third, the act of being able to evaluate teachers itself is likely to motivate students. There is a large literature showing that the opportunity to give feedback creates perceptions of fairness and procedural justice (Tyler & Caine, 1981). Research has shown that people believe systems are fairer when they have a voice, what Leventhal (1980) calls *participatory decision making* (see also Lind, Kanfer, & Earley, 1990). For example, workers believe that compensation schemes are fairer to the extent they have input into those schemes (Lawler, 1971).

Such evaluation opportunities also have motivational consequences; people are more motivated to complete a task when they are able to give feedback about the task. For example, in one study workers were tasked with filling out boring paperwork. Workers who were allowed to give feedback on the task completed more than 10% more paperwork than workers who didn't get to give feedback (Lind, Kanfer, & Earley, 1990). Another study investigated layoffs at a company, and found the single most important predictor of procedural justice was whether employees had been able to provide input about the layoffs, and that procedural justice predicted job satisfaction and reduced intentions to leave the company (Kernan & Hanges, 2002). This finding applies to systems beyond the workplace; indeed, it has been argued that one of the reasons that democracies have less corruption and more productive workforces than non-democracies is because the opportunity to have a say is built into the political system (Oppenheimer & Edwards, 2012).

Although there has been a great deal of research on procedural justice in law and organizational behavior, there has been limited work on applying the concept to course evaluations. The work that has been done has largely focused on the fact that students who believe that course grading procedures are fair tend to give higher course evaluations (Tata, 1999; Tyler & Caine, 1981). In

[1] A similar bias is the curse of knowledge (Camerer, Loewenstein, & Weber, 1989) in which people who know a fact have trouble getting into the mindset of those who don't. The existence of this bias suggests that teaching evaluations from faculty peers may suffer from similar problems as teaching evaluations from students. In both cases, it is very difficult to evaluate the extent of knowledge that has been gained.

[2] It is worth noting that these issues are exacerbated by the fact that many teaching evaluations don't explicitly ask students how much they have learned, but rather ask less specific questions such as "Overall, how would you rate the quality of this class?" or "Overall, how would you rate the quality of instruction?" Thus, in addition to the fact that students may not be able to evaluate the extent of their learning, they may also not realize that they are supposed to be doing so. The ambiguity of the questions allows students to weight what they value in the class, and students may not all weight learning as highly as faculty would like. As Carpenter et al. (2020) suggest, refining the questions so that they actually reflect student learning could be helpful.

fact, course evaluations are more closely tied to beliefs about how fair the grading was than what grade the student actually received (Tyler & Caine, 1981)[3] . There have, however, been several forays into the relationship between procedural justice more generally and student motivation and outcomes. For example, Chory-Assad and Paulsel (2004) found that student perceptions of procedural justice in the classroom led to higher student motivation, increased willingness to take further classes on the topic, deeper commitment to using course material after the course is over, and reduced aggression. Others have found that students are more likely to voluntarily follow the rules and policies of a course if they believe the procedures governing the creation and implementation of those rules are fair (Colquitt, 2001).

Given the large literature showing that the opportunity to provide feedback drives perceptions of procedural justice, and the fact that procedural justice drives a host of positive educational outcomes, there is good reason to think that offering students a chance to complete course evaluations can itself lead to beneficial consequences. While, as Carpenter et al. (2020) note, there are serious issues with interpreting the feedback that is provided, the fact that students get to provide feedback is probably a good thing for their educational experience.

Fourth, students' evaluations of teaching—even as limited and problematic as they can be—could nonetheless motivate teachers to improve the quality of their instruction. Knowing that one will be evaluated can lead people to greater effort, either due to social pressure (e.g., Kruglanski & Mayseless, 1987), or due to the knowledge that most administrators weigh these evaluations heavily in their assessments of teaching effectiveness. Indeed, in the classic paper by Stephen Kerr's (1975),"On the folly of rewarding for A while hoping for B," he lays out countless examples of how poorly designed reward structures create perverse incentives. One of the domains he highlights is university hiring and promotion. He argues that by focusing so much of the tenure decision at research universities on publications rather than on teaching, universities are ensuring that faculty neglect teaching in favor of prioritizing research. As such, Kerr implies that if universities are serious about their teaching mission, there should be more rewards for "good" teaching and punishment for "bad" teaching. The mere fact that faculty are being assessed by their students via teaching evaluations may lead them to be more responsive to student needs and devote more time to improving their pedagogy.

Unfortunately, as Carpenter et al. (2020) argue compellingly, standard measures of evaluating teaching are flawed. By rewarding teachers on the basis of these flawed instruments, universities may also incentivize behaviors that are undesirable, such as grade inflation or bribing students with chocolate[4] . In some sense, this is unavoidable. It is nearly impossible to construct a perfect incentive scheme. As Darley's Law (Darley, 2001) states: "The more any quantitative performance measure is used to determine an individual's rewards, the more subject it will be to corruption pressures and the more it will distort the action and thought patterns of those it is intended to monitor."

For example, by measuring outcomes using a pre-specified, standardized test, universities would incentivize teaching to that test (Popham, 2001). If teachers were instead evaluated on their students' performance in subsequent, higher-level classes, that would incentivize teachers to weed out weaker students (i.e., by creating policies that force less prepared students to drop the class, so that only the best students who are most likely to achieve in higher level classes will be on their record). In other words, any way (that we can see) of evaluating teaching could create negative externalities and incentivize problematic behaviors.

Developing effective output measures is made even more challenging by the fact that there are many factors that affect student learning, only one of which is teacher quality. Student learning may also be influenced by preparedness of the student, availability and quality of peer groups, and time spent on the material, among other things. A teacher can develop innovative and pedagogically sound problem sets that students don't complete. A teacher can develop an exceptional lesson plan, but fail to impart knowledge to students who have slept only two hours the night before the class and thus cannot focus on the material. Even if we could perfectly measure "quality of teaching" as a construct, the correlation between that construct and amount of student learning would not necessarily be especially high.

So what is to be done? One approach is to reward effort toward teaching improvement rather than student evaluations or other measures of student learning. The notion is that faculty document steps they have taken to improve their pedagogy, such as attending teaching conferences or workshops, revising syllabi or course materials, and developing novel courses. In this way, teachers can demonstrate that they are working to be more effective. That is, in the absence of reliable and valid output measures, we should perhaps focus more on input measures.

As an added bonus, even master teachers who might easily get excellent evaluations can improve—there is no ceiling to teaching quality—and this scheme incentivizes them to do so. It is, in a way, akin to the formative versus summative assessment distinction (Harlen & James, 1997): the assessment is itself a process for improvement rather than merely a measure of accomplishment.

Conclusion

Carpenter et al. (2020) have skillfully documented that course evaluations are not related to student learning. This finding is troubling, as many schools use evaluations as a primary means of determining teaching quality when assessing tenure and promotion cases. One reason for the lack of connection between students' evaluations and their actual learning may be an attribute substitution occurring during the evaluation: it is hard to directly quantify how much one has learned, and it is easier to bring to mind how much one has enjoyed the class. Therefore, the evaluators may use the sense of enjoyment they had in a class as a proxy for how much they have learned, even though these two factors may not be related or may even be

[3] Although, of course, students who receive better grades usually believe that the process for allocating grades was fair.

[4] The current authors are of mixed minds about whether bringing chocolate to class is actually a negative.

negatively correlated. Carpenter et al. (2020) provide a number of nuanced suggestions for resolving the issues with teaching evaluations that they raise in their article, and do not call for abandoning teaching evaluations. Nonetheless, it would be easy for readers to, upon learning about the validity challenges to teaching evaluations, decide that the practice should be abolished entirely. However, before we jettison these tools, it is worth considering whether they may be effective at measuring classroom dimensions other than learning, and the motivational roles that they may play above and beyond their evaluative role.

While course evaluations should be treated with caution, they may still provide useful qualitative feedback, predict which teachers are inspiring their students to future study, and give students a voice, which creates a sense of fairness and comes with a host of positive externalities. Nonetheless, measuring a teacher's effectiveness at promoting student learning is an extremely challenging task, and tying outcomes to rewards (e.g., tenure) creates corruption pressures on any measurement system. As such, while we believe there is still value in having students evaluate their teachers, we recommend moving toward formative rather than summative teaching evaluations for the purpose of promotion and tenure decisions. By evaluating the improvement efforts that teachers are engaging in, we create the right incentives for lifelong development as educators.

Statement of Author Contributions

Dr. Oppenheimer and Dr. Hargis each independently brainstormed ideas for inclusion in this manuscript. We then discussed those ideas, and outlined them together, and delegated sections for drafting. Dr. Oppenheimer then assembled the drafted elements into a manuscript and edited that manuscript. Dr. Hargis then edited the result. Several iterations of back and forth editing occurred, along with several conversations where disagreements were hashed out. Both authors approved the final version of the manuscript.

Keywords: Teaching evaluations; Metacognition; Fluency; Procedural justice; Motivation; Perverse Incentives; Formative vs. summative evaluation

References

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13*(3), 219–235.

Basow, S. A., & Martin, J. L. (2012). Bias in student evaluations. In M. E. Kite (Ed.), *Effective evaluation of teaching: A guide for faculty and administrators* (pp. 40–49). Washington, DC: Society for the Teaching of Psychology.

Bavisi, A., Madera, J. M., & Hebl, M. R. (2010). The effect of professor ethnicity and gender on student evaluations: judged before met. *Journal of Diversity in Higher Education*, *3*, 245–256.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society*, *2*, 59–68.

Camerer, C., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: an experimental analysis. *The Journal of Political Economy*, *97*(5), 1232–1254.

Carpenter, S. K., Witherby, A. E., & Tauber, S. K. (2020). On students' (mis)judgments of learning and teaching effectiveness. *Journal of Applied Research in Memory and Cognition*, *9*, 137–151.

Colquitt, J. A. (2001). On the dimensionality of organizational justice: a construct validation of a measure. *Journal of Applied Psychology*, *86*, 424–445.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly Journal of Economics*, *126*(4), 1593–1660.

Chory-Assad, R. M., & Paulsel, M. L. (2004). Classroom justice: Student aggression and resistance as reactions to perceived unfairness. *Communication Education*, *53*(3), 253–273.

Cuddy, A. J., Fiske, S. T., & Glick, P. (2007). The BIAS map: behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology*, *92*(4), 631.

Darley, J. M. (2001). The dynamics of authority in organizations and the unintended action consequences. In J. M. Darley, D. M. Messick, & T. R. Tyler (Eds.), *Social Influences on Ethical Behavior in Organizations* (pp. 37–52). Mahwah, NJ: L.A. Erlbaum Assoc.

Fischhoff, B. (1975). Hindsight is not equal to foresight: the effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, *1*(3), 288.

Frederick, S., & Loewenstein, G. (1999). Hedonic Adaptation. *Well-being: The Foundations of Hedonic Psychology*, 302–329.

Haggag, K., Patterson, R.W., Pope, N.G., & Feudo, A. (under review). Attribution Biases in Major Decisions: Evidence from the United States Military Academy.

Harlen, W., & James, M. (1997). Assessment and learning: differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice*, *4*(3), 365–379.

Kahneman, D., & Frederick, S. (2005). A model of intuitive judgment. In *The Cambridge handbook of thinking and reasoning*. pp. 267–293.

Kernan, M. C., & Hanges, P. J. (2002). Survivor reactions to reorganization: antecedents and consequences of procedural, interpersonal, and informational justice. *Journal of Applied Psychology*, *87*(5), 916–928.

Kerr, S. (1975). On the folly of rewarding A, while hoping for B. *Academy of Management Journal*, *18*, 769–783.

Krueger, A. B., & Whitmore, D. M. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: evidence from Project STAR. *The Economic Journal*, *111*(468), 1–28.

Kruglanski, A. W., & Mayseless, O. (1987). Motivational effects in the social comparison of opinions. *Journal of Personality and Social Psychology*, *53*(5), 834.

Lawler, E. E. (1971). *Pay and organizational effectiveness: A psychological view*. New York: Mcgraw Hill.

Leventhal, G. S. (1980). What should be done with equity theory? In K. J. Gergen, M. S. Greenberg, & R. H. Weiss (Eds.), *Social exchange: Advances in theory and research* (pp. 27–55). New York: Plenum.

Lind, E. A., Kanfer, R., & Earley, P. C. (1990). Voice, control, and procedural justice: instrumental and noninstrumental concerns in fairness judgments. *Journal of Personality and Social Psychology*, *59*, 952–959.

Oppenheimer, D. M., & Edwards, M. A. (2012). *Democracy despite itself: Why a system that shouldn't work at all works so well*. Cambridge: MIT Press.

Popham, W. J. (2001). Teaching to the test? *Educational leadership*, *58*(6), 16–21.

Protzko, J. (2015). The environment in raising early intelligence: a meta-analysis of the fadeout effect. *Intelligence*, *53*, 202–210.

Sanchez, C. A., & Khan, S. (2016). Instructor accents in online education and their effect on learning and attitudes. *Journal of Computer Assisted Learning*, *32*, 494–502.

Tata, J. (1999). Grade distributions, grading procedures, and students' evaluations of instructors: a justice perspective. *The Journal of Psychology*, *133*, 263–271.

Tyler, T. R., & Caine, A. (1981). The influence of outcomes and procedures on satisfaction with formal leaders. *Journal of Personality and Social Psychology*, *41*, 642–655.

Received 1 March 2020;
received in revised form 1 March 2020;
accepted 1 March 2020
Available online xxx