Journal of Applied Research in Memory and Cognition 9 (2020) 181–185

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

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

Commentary

On Students' (Mis)judgments of Learning and Teaching Effectiveness: Where We Stand and how to Move Forward

Shana K. Carpenter and Amber E. Witherby
Iowa State University, USA

Sarah K. Tauber
Texas Christian University, USA

Our target article addresses the complex and sometimes controversial topic of evaluating teaching, with a focused discussion of the factors that mislead students' judgments of their own learning and their evaluations of their teachers, and the problems associated with over-reliance on student evaluations in personnel decisions (Carpenter, Witherby, & Tauber, 2020). Six commentaries contributed many valuable insights on these issues, and we would like to express our thanks to the authors for taking the time to share their ideas and perspectives. They have raised exactly the kind of stimulating and thought-provoking discussions we had hoped for. Below we identify the primary themes emerging from the commentaries, and we offer our thoughts on these themes within the broader context of evaluating teaching.

The Multidimensional Nature of Teaching

A common thread running through the target article and commentaries is the fact that teaching is multidimensional. Though student evaluations of teaching provide information about students' experiences with a given course and instructor, they are but one measure that does not capture the various other components of effective teaching. Evaluating the complex and multidimensional nature of teaching is no small undertaking, and we agree with Boysen (2020) and Gurung (2020) about the importance of systematic efforts to formulate clear criteria for what it means to be an effective teacher. In particular, considerable efforts invested by psychological scientists have contributed to the identification of key behaviors of master teachers (Keeley, Furr, & Buskist, 2010), along with instructor attributes and practices that are relevant to effective teaching, such as an instructor's training, assessment practices, goals for student learning, and the instructional approaches that are used to accomplish those goals (Richmond et al., 2014). As any evaluation depends upon a clear understanding of what is being evaluated, defining the criteria for effective teaching is foundational to the design of a valid and reliable system to evaluate teaching, which undoubtedly requires multiple measures.

Boysen (2020) suggests that we overemphasize learning as a measure of teaching effectiveness. Though we certainly agree that multiple measures are needed, we hold that student learning should be key among these measures. The very idea of teaching somebody something entails some knowledge or skill that is gained as a result of that teaching. For example, students who are taught how to read in elementary school show evidence of that teaching in their abilities to sound out words and construct meaning out of written content, which they could not do before they were taught how to read. The knowledge or skills that are important at higher levels of education (e.g., critical thinking) may differ, but in the same way at all levels of education, a teacher facilitates students' learning.

Another reason to emphasize learning is because it can be objectively measured—not perfectly, of course, but more objectively than other measures of teaching effectiveness that are based on subjective impressions. Objective measures are particularly important given the vulnerability of subjective

Author Note.
Shana K. Carpenter & Amber E. Witherby, Department of Psychology, Iowa State University, USA.
Sarah K. Tauber, Department of Psychology, Texas Christian University, USA.

This material is based upon work supported by the James S. McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition, Collaborative Grant No. 220020483.

* Correspondence concerning this article should be addressed to Shana K. Carpenter, Department of Psychology, Iowa State University, W112 Lagomarcino Hall, 901 Stange Road, Ames, IA 50011, USA. Contact: shacarp@iastate.edu.

impressions to biases and misleading heuristics. In the target article, we discuss the ways in which highly intuitive factors (e.g., a fluent and enthusiastic teaching style) can increase students' subjective impressions of the effectiveness of their teachers, without having any effect on their objective learning. Subjective impressions can thus be misleading and hinder opportunities to accurately understand one's own learning, and in this way relying on those impressions to determine teaching effectiveness might be analogous to making financial decisions based on trying to guess the amount of money in one's bank account rather than directly verifying it.

The research we review primarily addresses direct measures of learning (e.g., students' grades in the current course and follow-up courses). However, several commentaries raise an important point that teachers might facilitate learning through indirect means as well. Even if a fluent or enthusiastic instructor does not directly affect students' content learning, does that instructor have other, non-immediate effects that could indirectly benefit learning? Oppenheimer and Hargis (2020) note that teachers can have positive downstream effects on students' academic behaviors. They offer some compelling examples, including a longitudinal analysis of the effects of students' kindergarten instruction on later educational outcomes (see Chetty et al., 2011). Though the quality of kindergarten instruction had diminishing effects on students' grades as they got older, it did predict outcomes much later in their lives, such as the likelihood of taking college entrance exams, enrolling in college, and future salary.

Furthermore, as suggested in multiple commentaries (Finn, 2020; Oppenheimer & Hargis, 2020; Serra & McNeely, 2020), teachers can influence students' enjoyment in a course and might stimulate their curiosity in a subject, both of which may influence students' decisions such as whether they will take another related course in the future. Some evidence supporting these possibilities may be found in some of our recent work (Carpenter, Northern, Tauber, & Toftness, 2020), showing that instructor fluency did not reliably affect objective learning, but did significantly increase several potential indirect measures of academic success, such as students' interest in the material and motivation to learn about it, as well as their likelihood of attending class, participating in class, studying the material, and applying the material to their lives. It is noteworthy that these are self-reported measures and not measures of actual student behavior, however, and here again we emphasize the importance of verifying the impact of instructor behaviors on indirect indices of learning, because those effects may not be intuitive or predictable. Longitudinal research on the indirect and downstream effects of instructor's approaches on learning-related outcomes is therefore of critical importance and should form the basis for the evidence supporting those approaches.

Data on these indirect and downstream effects could complement current measures of teaching effectiveness, while still emphasizing factors that promote learning. For example, along with the typical approach involving the collection of end-of-term student evaluations, students could be contacted at a future time and asked to provide information about how a given course and instructor influenced them over the longer-term. These follow-up assessments are most certainly subject to their own limitations (e.g., instructors of introductory courses would likely have the largest respondent pool), and ideally, they would be constructed from clearly defined goals about what long-term outcomes a given course is expected to produce. They may be challenging to implement, but such follow-up measures could provide insights into the important and often overlooked downstream effects of teaching. Interpreting outcomes from these measures would also need to be considered. Are indirect effects key for evaluating junior faculty? Are they instead an important metric for considering senior-level promotion or for identifying master teachers who may be excellent role models for junior faculty or for training students? It will be critical to develop clear plans for how to use and interpret the impact of indirect effects on student learning for personnel decisions in conjunction with other metrics of teaching effectiveness.

Leveraging Fluency for the Greater Good

In the target article, we review studies showing that students often prefer learning experiences that feel easier or more fluent, over experiences that feel subjectively more difficult. Learning can be undermined if students consistently choose learning strategies that feel easy but are actually less effective than those involving "desirable difficulties" (Bjork, Dunlosky, & Kornell, 2013), and the choice to avoid the latter may be reinforced by misleading affective cues associated with effort that can be a false signal of failure to learn (Kirk-Johnson, Galla, & Frauendorf, 2019). A key challenge, therefore, is getting students to engage with effective learning techniques that may be difficult and feel undesirable for them to adopt.

Here is where fluency could be leveraged. Finn (2020) describes a line of research on the benefits of "remembered success" in learning situations, wherein students' memory of success amidst a difficult learning experience can enhance their motivation and approach to future learning events. For instance, in a challenging learning situation (e.g., solving difficult math problems), inserting a less aversive version of the task at the end of the experience (e.g., math problems that were much easier) leads students to prefer the task and be more likely to engage with it again in the future relative to a task that ends with the challenging problems (Finn, 2010; Finn & Miele, 2016; for a review see Finn, 2015). In this way, fluency during learning need not be avoided and could actually be quite useful. Similar to the spoonful of sugar that mollifies the bad-tasting medicine, a dose of fluency as part of a challenging learning experience may give students the motivation they need to persist, and to pursue further challenging learning experiences.

The positive impression that students have of fluent, engaging, or enthusiastic instructors may be leveraged in other ways as well. Although classroom-based research shows that such factors do not improve students' course performance (Bettencourt, Gillett, Gall, & Hull, 1983; Serra & Magreehan, 2016; Williams & Ceci, 1997), these factors may have positive effects on students' impressions of teachers that might be farther-reaching than we knew. Serra and McNeely (2020) report new data from a classroom study showing that instructors who are rated as

more fluent are not only rated by students as more effective teachers (after controlling for course grades), but also rated by students as more likely to secretly be Batman, or to be capable of choreographing Beyoncé's music videos. Though most instructors are likely not capable of these things (we leave it up to individual instructors to preserve or dispel this illusion for their own students), Serra and McNeely note that this finding might reflect a halo effect that raises the interesting possibility that well-liked instructors may have the power to positively influence students' academic behaviors. That is, if students have faith in their teacher's super powers and dance choreography skills (or if such measures are simply a proxy for some other positive attributes, such as general likeability), students may also be more likely to listen to their teacher's advice. On the important assumption that such advice involves evidence-based recommendations about how to study and learn effectively, a well-regarded instructor could facilitate students' tendencies to develop effective study habits. This is another interesting possibility that could be empirically investigated.

The Utility of Student Evaluations of Teaching

In our target article, we review the results of studies that raise questions about the validity of student evaluations of teaching as indicators of learning, based on a number of documented biases and a lack of a positive relationship between student ratings of teaching effectiveness and objective measures of learning. A number of commentaries resonated with this notion, and assert that student evaluations of teaching primarily reflect students' satisfaction with their courses and instructors (Boysen, 2020; Finn, 2020; Kornell, 2020).

Satisfaction ratings fluctuate and can be sensitive to biases, and it is important to consider them in the context of curriculum (e.g., whether the course is required, or whether there is a history of high ratings for the course, regardless of who teaches it, because students enjoy the content). Further, Kornell (2020) notes that student ratings, especially when persistently and chronically low, can be effective for revealing problems with a course or instructor. This is an important point, and it is worth highlighting the fact that even the phantom professors received ratings in the range of average to above average (Reynolds, 1977; Uijtdehaage & O'Neal, 2015). This suggests that extremely low ratings may be reflective of a course experience that is worse than no course experience at all. To the extent that these low ratings reflect an instructor's failure to carry out reasonable responsibilities (e.g., showing up for office hours, answering students' questions), we hope all educators would agree that students' satisfaction with their courses should reflect instructional practices that are better than nothing.

There are multiple ways to receive low ratings. Indeed, the fact that there are multiple routes to the same place renders it impossible to infer the cause from the outcome. Although certain factors that are good for learning can deflate students' ratings of instructors (e.g., active learning, Deslauriers, McCarty, Miller, Callaghan, & Kestin, 2019), an inverse interpretation of this relationship might suggest that instructors who receive low ratings are doing good things for students' learning. This is not always true, of course, but the ratings alone can neither verify nor falsify this, and therein lies the problem with relying heavily on students' ratings as a measure of teaching effectiveness.

A more general notion expressed in the commentaries is the idea that being evaluated could maintain quality assurance and provide motivation for instructors to improve (Boysen, 2020; Finn, 2020; Kornell, 2020; Oppenheimer & Hargis, 2020). What to improve upon depends on the goals of teaching, and we emphasize that student feedback should be carefully interpreted in light of the fact that sometimes goals for learning may not coincide with goals of pleasing students. For example, giving students stickers because they like them is more aligned with goals for increasing students' enjoyment than with facilitating their learning (Boysen, 2020). Goal-driven approaches to facilitate learning might involve things that students do not always enjoy (for example, quizzes), but foregoing such approaches because they are disliked by students—something that instructors may find themselves tempted to do if the goal is to increase satisfaction ratings—undermines the mission of teachers to serve as facilitators of learning. When the goal is learning, student feedback can be valuable, particularly if it helps an instructor improve upon the quizzing methods in a way that more effectively facilitates learning, but such feedback for the sake of teaching improvements again should be carefully interpreted in light of how it helps facilitate learning goals.

The problem that we identify in the target article is, of course, that students' ratings can be influenced by things that have no effect on their learning, including instructor race, gender, age, and accent. It therefore appears that baseline ratings already exist for different instructors, regardless of their teaching effectiveness. Some instructors are at a disadvantage regardless of what happens in their classes, which is a serious problem. Students' ratings can also be inflated by lenient grading and the withholding of challenging learning experiences. In this way, based on student ratings alone it can be hard to tell the difference between an instructor who invests tremendous time and energy into shepherding students through the challenges, uncertainties, and frustrations involved in intellectual growth, compared to an instructor who foregoes these time commitments but hands out high grades. Increased demands on instructors' time, combined with heavy reliance on student ratings in personnel decisions, create incentives for instructors to take time-saving shortcuts that may boost ratings but not provide the best educational experience for students. We do not argue, as suggested by Boysen (2020), that this incentive is created by the existence of student evaluations, but rather in the way that those evaluations can be misinterpreted and in how easily they can be gamed.

Nor do we suggest that the baby should be thrown out with the bathwater. Instructors who are well-liked can use effective learning techniques and hold students to high academic standards. Well-liked instructors may even catalyze students' uptake of effective learning strategies (Serra & McNeely, 2020). As Gurung (2020) notes, such change is neither easy nor immediate, with metacognitive calibration issues persisting well past the "reality check" from the first exam. Indeed, adoption of effective learning strategies has been shown to occur under conditions in which the target behaviors (e.g., using empirically-validated

study strategies) are reinforced by instructors who invest the time to explicitly teach students about these strategies, demonstrate them, remind students about them, and show students the outcomes of using such strategies (e.g., see Biwer, oude Egbrink, Aalten, & de Bruin, 2020; Carpenter et al., 2017). Such an approach takes a patient and dedicated teacher who, unfortunately, may not be recognized by a system that relies too heavily upon students' satisfaction ratings as primary measures of teaching effectiveness.

Implications for Evaluating Teaching

What then can be said about the best ways to evaluate teaching? Few would disagree that we need multiple measures. In addition to the alternatives and supplements to student evaluations that we review in the target article, the commentary writers offer some excellent suggestions such as soliciting students' self-reports of effective pedagogical approaches (e.g., distributed practice) that were used in their courses (Boysen, 2020), and having instructors keep documented evidence of their efforts to improve their teaching (Oppenheimer & Hargis, 2020). As with all measures of teaching effectiveness, we propose that these be coupled with evidence of their relation to improved student learning. More evidence-based measures of teaching effectiveness are superior to fewer measures, and it is critical to develop detailed strategies for how outcomes will be interpreted for meeting the educational goals of the instructor, department, and institution.

Controversy over evaluating teaching will undoubtedly continue. There is currently no single measure free from biases, and new supplemental and alternative approaches may be proposed in the years to come. Looking ahead, we emphasize the importance of evidence-guided criteria for evaluating teaching. Sound research on the factors that enhance student learning, motivation, inspiration, and other academic behaviors can offer a strong basis for guiding specific instructional approaches that are designed to achieve those goals. Collecting data from the relevant measures (student feedback, course performance, follow-up assessments) that reflect attainment of those goals can then be used to make necessary improvements to our approaches and further contribute to the evidence guiding those decisions. Grounding our teaching approaches in empirically-validated research carries with it a number of benefits, including (a) encouraging us to be knowledgeable about the effects of the pedagogical practices we are using, (b) inoculating us from the many intuitive yet misleading illusions of learning that abound in education, and (c) documenting "walk the walk" efforts that can distinguish a committed teacher from one who takes the all-too-easy shortcuts to inflate students' ratings.

Conclusions

Students' ratings of their teachers can be useful for identifying problems with a course and providing feedback about students' experiences, but these ratings are subject to bias and are often over relied-upon as measures of teaching quality. Effective teaching is multifaceted and likely has long-term influences on students' academic behaviors and decisions that go unmeasured. Collecting data on such things is possible, but these data are complex, time-consuming to collect, and not easily boiled down to a single number. As a consequence, they probably introduce inconveniences that make them difficult or undesirable to adopt in personnel decisions. Desirable difficulties can be beneficial at all levels of education, however, and we encourage everyone who is invested in high-quality teaching to take the time to carefully and thoroughly evaluate the efforts that good teachers are investing into this important mission.

Author Contributions

All three authors conceived the content of this reply, contributed to the writing, and provided edits prior to submission.

Keywords: Learning, Metacognition, Education, Teaching evaluations

References

Bettencourt, E. M., Gillett, M. H., Gall, M. D., & Hull, R. E. (1983). Effects of teacher enthusiasm training on student on-task behavior and achievement. *American Educational Research Journal*, *20*, 435–450.

Biwer, F., oude Egbrink, M. G. A., Aalten, P., & de Bruin, A. B. H. (2020). Fostering effective learning strategies in higher education: A mixed methods study. *Journal of Applied Research in Memory & Cognition*, *9*, 186–203.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*, 417–444.

Boysen, G. A. (2020). The multidimensional nature of teaching and student evaluations: Commentary on students' judgements of learning and teaching effectiveness. *Journal of Applied Research in Memory and Cognition*, *9*, 152–156.

Carpenter, S. K., Northern, P. E., Tauber, S. K., & Toftness, A. R. (2020). Effects of lecture fluency and instructor experience on students' judgments of learning, test scores, and evaluations of instructors. *Journal of Experimental Psychology: Applied*, *26*, 26–39.

Carpenter, S. K., Rahman, S., Lund, T. J. S., Armstrong, P. I., Lamm, M. H., Reason, R. D., & Coffman, C. R. (2017). Students' use of optional online reviews and their relationship to summative assessment outcomes in introductory biology. *CBE Life Sciences Education*, *16*, 1–9.

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.

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*, 1593–1660.

Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., & Kestin, G. (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.

Finn, B. (2010). Ending on a high note: Adding a better end to difficult study. *Journal of Experimental Psychology: Learning Memory, and Cognition*, *36*, 1548–1553.

Finn, B. (2015). Retrospective utility of educational experiences: Converging research from education and judgment and decision-making. *Journal of Applied Research in Memory & Cognition*, *4*, 374–380.

Finn, B. (2020). What more can we learn from teaching evaluations? *Journal of Applied Research in Memory and Cognition*, *9*, 157–160.

Finn, B., & Miele, D. B. (2016). Hitting a high note on math tests: Remembered success influences test preferences. *Journal of Experimental Psychology: Learning Memory, and Cognition*, *42*, 17–38.

Gurung, R. A. R. (2020). Call it out: Recognizing good teaching and learning. *Journal of Applied Research in Memory and Cognition*, *9*, 161–164.

Keeley, J., Furr, R. M., & Buskist, W. (2010). Differentiating psychology students' perceptions of teachers using the teacher behavior checklist. *Teaching of Psychology*, *37*, 16–20.

Kirk-Johnson, A., Galla, B. M., & Frauendorf, S. H. (2019). Perceived effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice. *Cognitive Psychology*, *115*, 1–31.

Kornell, N. (2020). Why and how you should read student evaluations of teaching. *Journal of Applied Research in Memory and Cognition*, *9*, 165–169.

Oppenheimer, D. M., & Hargis, M. B. (2020). If teaching evaluations don't measure learning, what do they do? *Journal of Applied Research in Memory and Cognition*, *9*, 170–174.

Reynolds, D. V. (1977). Students who haven't seen a film on sexuality and communication prefer it to a lecture on the history of psychology they haven't heard: Some implications for the university. *Teaching of Psychology*, *4*, 82–83.

Richmond, A. S., Boysen, G. A., Gurung, R. A. R., Tazeau, Y. N., Meyers, S. A., & Sciutto, M. J. (2014). Aspirational model teaching criteria for psychology. *Teaching of Psychology*, *41*, 281–295.

Serra, M. J., & Magreehan, D. A. (2016). Instructor fluency correlates with students' ratings of their learning and their instructor in an actual course. *Creative Education*, *7*, 1154–1165.

Serra, M. J., & McNeely, D. A. (2020). The most fluent instructors might choreograph for Beyoncé or secretly be Batman: Commentary on Carpenter, Witherby, and Tauber. *Journal of Applied Research in Memory and Cognition*, *9*, 175–180.

Uijtdehaage, S., & O'Neal, C. (2015). A curious case of the phantom professor: Mindless teaching evaluations by medical students. *Medical Education*, *49*, 928–932.

Williams, W. M., & Ceci, S. J. (1997). "How'm I doing?" Problems with student ratings of instructors and courses. *Change: The Magazine of Higher Learning*, *29*, 12–23.

Received 11 April 2020;
accepted 12 April 2020
Available online xxx