Journal of Experimental Psychology: Applied

1076-898X/16/$12.00 http://dx.doi.org/10.1037/xap0000077

The Effect of Instructor Fluency on Students' Perceptions of Instructors, Confidence in Learning, and Actual Learning

Shana K. Carpenter
Iowa State University

Laura Mickes
Royal Holloway, University of London

Shuhebur Rahman and Chad Fernandez
Iowa State University

Students' judgments of their own learning are often misled by perceptions of *fluency*—the ease with which information is presented during learning. Lectures represent important learning experiences that contain variations in fluency, but have not been extensively studied. In the current study, students watched a 22-min videotaped lecture that was delivered by the same instructor in either a *fluent* (strong, confident, and deliberate) manner, or in a *disfluent* (uncertain, hesitant, and disengaged) manner. Students then predicted their score on an upcoming test on the information, rated the instructor on traditional evaluation measures, and took a multiple-choice test on the information immediately (Experiment 1), after 10 min (Experiment 2), or after 1 day (Experiment 3). The fluent instructor was rated significantly higher than the disfluent instructor, but test scores did not consistently differ between the 2 conditions. Though students did not indicate higher confidence overall in learning from a fluent instructor, Experiment 3 found that when participants base their confidence on the instructor, those in the fluent condition were more likely to be overconfident. These findings indicate that instructor fluency leads to higher ratings of instructors and can lead to higher confidence, but it does not necessarily lead to better learning.

Keywords: memory, metacognition, fluency, instructor evaluations

The path to successful learning requires students to accurately evaluate their own knowledge. Students' impressions of how well they understand a concept can influence their study decisions, and as a consequence, their performance on course-related assessments. With advances in technology that afford more educational opportunities outside of traditional classroom settings, it is becoming increasingly important for students to effectively monitor and regulate their own learning.

Unfortunately, there is often a gap between students' impressions of how much they know about something and the objective verification—via a test or assignment—of how much they really know. Decades of research on metacognition has shown that students tend to overestimate their own knowledge. When asked to predict their own performance on an upcoming test, the predictions that students give are often higher than their actual performance on the test. This has been shown in many laboratory studies (e.g., Castel, McCabe, & Roediger, 2007; Dunlosky & Metcalfe, 2009; Dunlosky & Nelson, 1994; Finn & Metcalfe, 2007; Koriat & Bjork, 2005; Koriat, Sheffer, & Ma'ayan, 2002; Kornell & Bjork, 2009), and also in classroom studies where students often overpredict their performance on upcoming assessments over course material that they are currently learning (e.g., Bol, Hacker, O'Shea, & Allen, 2005; Carpenter et al., 2015; Hacker, Bol, Horgan, & Rakow, 2000; Miller & Geraci, 2011).

In academic situations, overconfidence can lead to the unfortunate and sometimes surprising realization that students experience when they are confronted with the fact that they have performed worse than they expected. The negative consequences of overconfidence can be difficult to overcome. Even if students' metacognitive awareness improves with practice (i.e., the "reality check" they get after the first exam) and their scores improve on subsequent exams, one low exam score can account for a nontrivial portion of their final course grade. The subjective experience of low performance can also be accompanied by other undesirable consequences, such as academic disengagement and attrition (Baillie & Fitzgerald, 2000; Geisinger & Raman, 2013). Thus, understanding the factors that contribute to overconfidence, and how they might apply in academic situations, is critical to improving students' success and persistence.

Shana K. Carpenter, Department of Psychology, Iowa State University; Laura Mickes, Department of Psychology, Royal Holloway, University of London; Shuhebur Rahman and Chad S. Fernandez, Department of Psychology, Iowa State University.

This material is based upon work supported by the National Science Foundation under Grant DUE-1504480. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank Rosa Cai, Rachel Conn, Brianna Culich, Torie Homervold, Amanda Malin, Vivian Spilde, and Rachel Teague for their assistance with data collection and scoring.

Correspondence concerning this article should be addressed to Shana K. Carpenter, Department of Psychology, Iowa State University, W112 Lagomarcino Hall, Ames, IA 50011. E-mail: shacarp@iastate.edu

Research on metacognition has revealed that overconfidence arises when students base their judgments of learning on factors that are not diagnostic of their actual learning. Whereas some factors can be reliable indicators of a student's level of knowledge (e.g., one's performance on a practice assessment), other factors are poor indicators and can even be inversely related to a student's level of knowledge. One of the most widely studied factors that can mislead students' perceptions of their own learning is *fluency*, or the perceived ease with which information is processed during learning (for recent reviews, see Alter & Oppenheimer, 2009; Bjork, Dunlosky, & Kornell, 2013; Finn & Tauber, 2015). Some studies have shown, for example, that students' predictions of their own performance on an upcoming test are higher—but performance itself is not higher—for information that is presented to them in an easier-to-read font style (Alter, Oppenheimer, Epley, & Eyre, 2007), or in a larger font size (Rhodes & Castel, 2008).

Other studies have shown that students' predictions of performance—but not actual performance—are higher when verbal information is accompanied by colorful images and graphics, such as pictures appearing alongside text descriptions of scientific phenomena (Serra & Dunlosky, 2010), or pictures denoting the English translations of foreign language vocabulary words (Carpenter & Olson, 2012). In these studies the presence of a picture, although it did not benefit memory for the verbal information that it accompanied, created an impression that the material was easier to process and thus would be easier to remember. Direct evidence for this ease-of-processing heuristic comes from Carpenter and Olson's (2012) Experiment 4, in which participants were given unfamiliar foreign language words—either accompanied by pictures denoting their meaning, or by their English translations—and asked to rate how easy it was to study the pair of items together, how easy it was to understand the foreign word from the picture (vs. the English translation), and how easy it was to link the meaning of the foreign word to the picture (vs. the English translation). In all cases, participants' ease-of-processing ratings were higher when the foreign words were accompanied by pictures than by English translations.

The illusion of fluency can also lead students to misjudge the effectiveness of different learning techniques. Many studies have demonstrated the reliable and powerful benefits of spaced practice—repeatedly studying information in a way that is distributed across time rather than massed together in immediate repetition (for recent reviews, see Carpenter, Cepeda, Rohrer, Kang, & Pashler, 2012; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008; Delaney, Verkoeijen, & Spirgel, 2010; Gerbier & Toppino, 2015; Küpper-Tetzel, 2014). However, students often feel more confident in their learning following massed practice compared to spaced practice (Carpenter & Mueller, 2013; Kornell & Bjork, 2007; Simon & Bjork, 2001). Even after having a chance to experience both techniques and demonstrating greater objective learning from spaced versus massed practice, students still adopt the erroneous belief that massed practice was more effective in helping them learn (Kornell & Bjork, 2008).

Students' tendency to endorse massed practice could arise from the sense of fluency that it provides. When material is repeatedly encountered in immediate succession, it is readily available in short-term memory, creating the impression that it has been well-learned. The ease with which information comes to mind in the short-term, however, is not always a good indicator of long-term learning. Though recalling information on tests that occur at massed repetitions is much easier initially and leads to higher accuracy than recalling information on tests that occur at spaced repetitions, this pattern reverses in the long term, such that memory assessed after a delay reveals an advantage for information learned via spaced tests (Carpenter & DeLosh, 2005; Carpenter, Pashler, & Cepeda, 2009). Results like these reveal an important distinction between the perceived ease of processing during initial learning and the durability of long-term retention, which can sometimes be inversely related.

Academic settings afford the opportunity for students to be vulnerable to this "illusion of knowing" driven by fluency. Students routinely encounter information that varies in its perceived ease of processing. In particular, in any college or university there is wide variation in instructors' teaching styles. Some instructors, due perhaps to years of experience, give smooth and well-polished lectures, whereas others are less organized and may fumble through the more difficult parts. The appearance of how easy the information is to learn—based on the ease with which the instructor explains it—may influence students' judgments of how easy it will be for them to remember. Lecture-based learning is one area where variations in ease of processing abound, but the effects they might have on students' confidence and learning are currently not well understood.

One recent study explored this by manipulating the fluency of a lecture and its effects on students' perceived and actual learning. Carpenter, Wilford, Kornell, and Mullaney (2013) had students watch one of two prerecorded lecture videos of an instructor explaining a scientific concept. The same instructor appeared in both videos, and the content taught was scripted to ensure that it was identical across the two videos. The only difference between the two videos was in how the instructor delivered the lecture. In the *fluent* condition, the instructor stood facing the camera, explaining the material in a confident and fluid manner without help from notes. In the *disfluent* condition, she delivered the same content while hunched over a desk, reading from notes, stumbling over words and pausing awkwardly.

After watching one of these two videos students rated the instructor on traditional teacher evaluation measures, including preparedness, organization, knowledge, and overall effectiveness. The fluent instructor received average ratings that were significantly higher than the disfluent instructor (4.2 vs. 1.5 on a 5-point scale), however a later memory test revealed no significant difference in learning between the two conditions. This was true even though students in the fluent condition estimated their knowledge of the material to be significantly higher than those in the disfluent condition. More specifically, when asked to predict their future test performance immediately after watching the video, students in the disfluent condition predicted a level of performance that was close to what they actually attained on the memory test. Students in the fluent condition, on the other hand, predicted that they would recall about twice as much as they actually did.

This study provides some evidence that the misleading effects of fluency might apply to lecture-based learning. This carries important implications for designing lectures in a way that is most effective for student learning and helps them avoid the pitfalls of overconfidence. Many handbooks on college teaching encourage instructors to prepare well-organized and engaging lectures (e.g., Brown & Atkins, 1990; Brown & Race, 2002; Davis, 1993; Ekeler,

1994; Hogan, 1999; Lowman, 1995; Morton, 2009). Though it seems highly intuitive that students would learn better from a smooth and well-polished lecture compared to a fumbled and awkward one, data from the Carpenter et al. (2013) study suggest that this may not be the case, and that a well-polished and engaging lecture could even introduce the undesirable effect of overconfidence.

There are some notable limitations to the Carpenter et al. (2013) study, however. First, the lecture videos were quite brief, lasting only about 1 minute. It is possible that fluent instructors do enhance learning—perhaps by encouraging student engagement and discouraging boredom or "zoning out," during class—but this effect did not occur in the study because the videos were so brief. If students can maintain their attention to a video for 1 minute, they may not have a chance to fall prey to the boredom and disengagement that could result from a longer disfluent lecture of the length that is typically encountered in an actual class.

Second, the amount of information to be learned consisted of only 10 relatively simple idea units. Materials encountered in a class are often more complex, consisting of several key concepts and connections between those concepts. With such materials, a disfluent instructor may create distraction or confusion that interferes with students' ability to concentrate and make connections between concepts, leading to negative effects on memory retention. Thus, there are reasons to suspect that with longer and more complex materials, fluent lectures might lead to better learning than disfluent lectures.

There are also reasons to suspect that students' overconfidence in their own learning may not be as high as it was found to be in Carpenter et al. (2013). In that study, the instructor was the sole source of the information presented. She appeared in the center of the videos and explained the content without the use of any visual aids, increasing the chances that students' attention would be focused on her. In actual courses, however, instructors use a variety of visual aids, technology, and other teaching materials to present the information, drawing students' attention away from the instructor's behavior and giving them a variety of additional information on which to base their judgments of learning. Under these conditions, students may not give as much weight to the instructor's behavioral tendencies in estimating how much they have learned, and thus may not be as likely to base their own confidence on the instructor's apparent confidence in the material.

Thus, under conditions that are more educationally realistic, there are reasons to expect that students may learn better from a fluent instructor than from a disfluent instructor, and that a fluent instructor may not lead to an illusion of knowing. The current study set out to test these predictions. Students viewed a 22-min lecture on signal detection theory that had been prepared for an actual class. The video was an animated presentation complete with graphics (see Figure 1). The video was accompanied by the voice of an instructor explaining the content, but the instructor was not seen in the video. The manipulation was carried out by modifying only the instructor's voice. Consistent with terminology used by Carpenter et al. (2013), we refer to these conditions as the fluent and disfluent conditions. In the fluent condition, the instructor spoke with confidence in a calm, smooth, and fluid tone throughout the video. In the disfluent condition, the instructor stumbled over words, paused awkwardly, made frequent use of "ums," and periodically trailed off while speaking. The visual content of the lecture—slides, animations, and timing—was identical in both videos, and the auditory content was scripted to be sure that the instructor presented the same information in both videos. The only difference between the two videos was the instructor's delivery—deliberate and confident (fluent), or hesitant and disengaged (disfluent).[1]

If students perceive a difference in the fluency of the instructor, then higher ratings of instructor behaviors reflecting fluency (organization, knowledge, preparedness, and effectiveness) should be assigned to the fluent instructor compared to the disfluent instructor. If fluency of instruction benefits learning of educationally realistic materials, then the positive relationship between instructor fluency and student learning outcomes should be apparent. Finally, if the lecture materials provide a variety of cues upon which students can base their confidence that are not restricted to the fluency of the instructor, students may be less likely to fall prey to an "illusion of knowing" and not readily demonstrate overconfidence in their own learning.

Experiment 1

Participants

Seventy-four participants were recruited from introductory-level psychology courses at Iowa State University, and received partial course credit in exchange for completing the study.

Design and Procedure

After giving informed consent to participate in the study, each participant was seated at a computer and asked to put on a pair of headphones. Instructions on the computer screen informed participants that they would be watching a video (approximately 20 min in length) of an instructor explaining a scientific concept, and that later their memory for the information in the video would be tested. Participants were not encouraged to take notes during the video.

After reading the instructions, participants began the experiment by viewing the video with either the fluent instructor ($n = 37$) or the disfluent instructor ($n = 37$). Immediately after the video ended, participants were instructed that they could remove the headphones. At that time, the following question appeared on the computer screen: "In about 1 minute from now we will give you a multiple-choice test on the information from the video. How well

[1] The term *fluency* has been used in the literature on metacognition to refer to the experienced ease of processing of stimuli during learning. This experience can be directly measured through the speed of participants' responses to particular stimuli during learning (e.g., Mueller, Tauber, & Dunlosky, 2013), or through participants' ratings of the ease of processing stimuli during learning (e.g., Carpenter & Olson, 2012). In the current study, we use the term fluency to refer to the behaviors of an instructor that reflect the smoothness of delivery of the lecture, consistent with the term as used in previous studies of lecture-based learning (Carpenter et al., 2013). Thus, fluency—as referred to here—does not refer to a response measure reflecting the manner in which participants overtly process the stimuli, but rather the lecture delivery style of the instructor. To the extent that such response measures are unaffected by lecture delivery style, it is likely that any effects of lecture delivery style on students' perceived learning are reflective of their preexisting beliefs about learning—that is, more competent instructors give more fluent lectures.

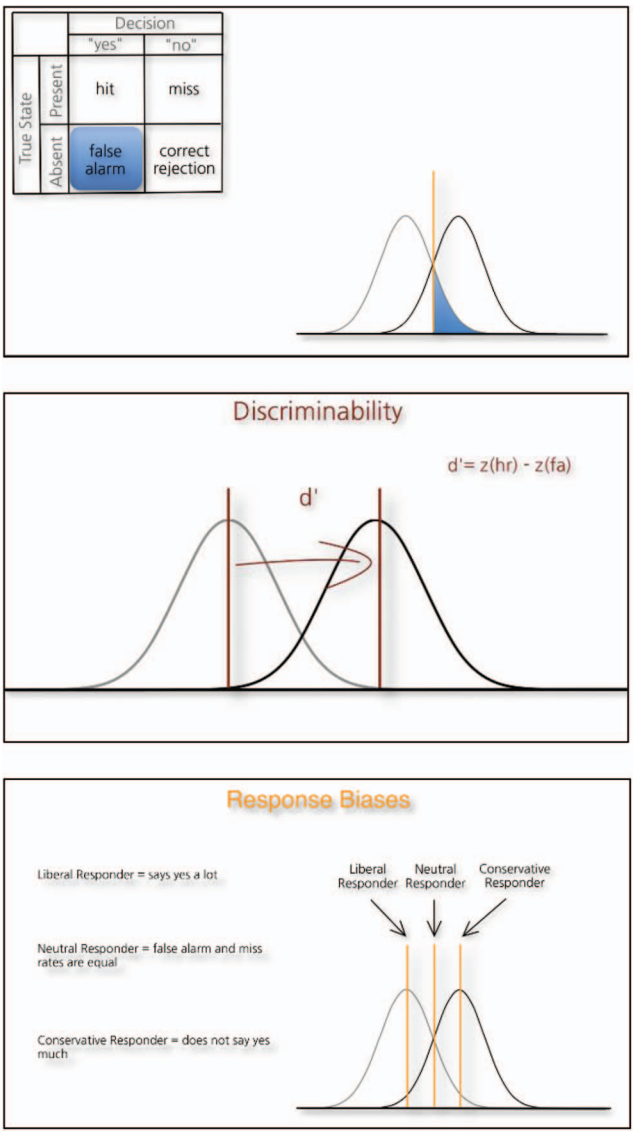

Figure 1. Screenshots of the lecture on signal detection theory. See the online article for the color version of this figure.

do you think you will score?" Participants were instructed to enter a value between 0% and 100%. Right after answering this question, participants completed instructor evaluation questions requiring them to rate (from 1–5) how knowledgeable, organized, and prepared the instructor was, followed by a 1–5 rating of the overall effectiveness of the instructor. Participants then completed questions requiring them to rate (on the same 1–5 scale) their own motivation and interest to learn the material, in addition to how well they felt they had learned the material. On each of these eight questions (the judgment of their own learning [JOL] followed by the seven evaluation questions), responses were self-paced.

Immediately after answering the last evaluation question, participants were given a 20-item multiple-choice test on the topics covered in the video. These questions consisted of relatively straightforward factual content from the video (e.g., "When something is present in the environment but the individual incorrectly says that it is not present, that is a: (a) hit, (b) miss, (c) false alarm, (d) correct rejection, (e) I don't know"). Each question included four alternatives (with only one being correct), and included an option to indicate "I don't know." The 20 questions appeared in a fixed order corresponding to the order in which the content appeared in the video. Participants answered one question at a time, and had unlimited time to answer each question.

After answering the last test question, participants answered one final question inquiring about whether they had any prior knowledge of signal detection theory before participating in the study. After answering this question participants were debriefed and thanked. Four participants reported having prior knowledge of signal detection theory, so their data were excluded from all analyses.

Results

Instructor evaluation ratings. Participants' mean ratings on the instructor evaluation questions are given in Table 1. The fluent instructor was rated significantly higher than the disfluent instructor on organization, $t(68) = 4.78$, $p < .001$, $d = 1.14$, knowledge, $t(68) = 3.07$, $p = .003$, $d = .72$, preparedness, $t(68) = 5.22$, $p < .001$, $d = 1.25$, and overall effectiveness, $t(68) = 5.29$, $p < .001$, $d = 1.26$. No significant differences were observed between the fluent and disfluent conditions in students' ratings of motivation (2.36 vs. 2.38, respectively), interest (2.11 vs. 2.00, respectively), or in how much they felt they had learned the material (2.78 vs. 2.62, respectively), ts <1.

Predicted versus actual performance. Scores on the multiple-choice test revealed no significant difference in student learning between the fluent condition ($M = 66\%$, $SD = 19\%$) and the disfluent condition ($M = 65\%$, $SD = 18\%$), $t(68) = .25$. Response times associated with correct responses also did not differ between the fluent condition ($M = 11.38$ s, $SD = 4.21$ s) and the disfluent condition ($M = 10.21$ s, $SD = 2.91$ s), $t(68) = 1.34$, $p = .18$. Overall judgments of learning were similar for the fluent condition ($M = 62\%$, $SD = 18\%$) compared to the disfluent condition ($M = 62\%$, $SD = 17\%$), $t(68) = .05$, indicating that students' confidence in their own learning was not significantly affected by instructor fluency. The correlation between students' JOLs and test scores was positive in both the fluent condition, $r = .40$, $p = .015$ and in the disfluent condition, $r = .34$, $p = .048$, indicating fairly consistent agreement between students' perceived learning and actual learning.

Discussion

These results indicate that the vocal cues of an instructor are sufficient to produce differences in students' perceptions of instructors based on fluency. However, a fluent instructor rated by students to be high in knowledge, preparedness, organization and effectiveness did not produce better learning than a disfluent instructor who was rated significantly lower on all of these measures. This is consistent with the findings reported by Carpenter et al. (2013), and inconsistent with our prediction that these educationally relevant materials might be more likely to yield a benefit in test scores for the fluent condition over the disfluent condition.

We note, however, that the multiple-choice test occurred immediately after the learning phase in Experiment 1. It is possible that

Table 1
Mean Instructor Ratings (1–5) Across the Fluent and Disfluent Conditions

	Organized	Knowledgeable	Prepared	Effective
Experiment 1 (Immediate test)				
Fluent condition	4.03 (.85)	4.69 (.47)	4.56 (.61)	3.72 (.85)
Disfluent condition	3.03 (.90)	4.15 (.96)	3.32 (1.27)	2.56 (.99)
Experiment 2 (10-Minute delayed test)				
Fluent condition	4.02 (.84)	4.53 (.64)	4.34 (.85)	3.09 (1.15)
Disfluent condition	3.16 (1.03)	4.00 (.76)	3.27 (1.08)	2.49 (1.10)
Experiment 3 (24-Hour delayed test)				
Fluent condition	4.25 (.72)	4.70 (.46)	4.68 (.56)	3.82 (.95)
Disfluent condition	3.28 (.90)	4.34 (.73)	3.70 (1.14)	2.70 (.99)

Note. Standard deviations are given in parentheses.

students forget information faster after a disfluent lecture compared to a fluent lecture, but this difference failed to emerge on an immediate test that allowed students to remember the information from relatively short-term memory. Experiment 2 was designed to provide conditions under which forgetting of the material was more likely to occur. Unlike in Experiment 1 where the test was provided immediately after learning, the test in Experiment 2 was provided after a 10-min delay.

Experiment 2

Participants

One hundred and four participants were recruited from the same participant pool as in Experiment 1. None of them had participated in Experiment 1.

Design and Procedure

Participants were randomly assigned to view either the fluent video ($n = 53$) or the disfluent video ($n = 51$), then made a JOL concerning how well they would score on a multiple-choice test given on the information after about 10 min. Participants then answered the same instructor evaluation questions from Experiment 1, and after a time interval lasting approximately 10 min that involved answering random trivia questions, completed the same 20-item multiple-choice test from Experiment 1, and then were fully debriefed. Two participants reported having prior knowledge of signal detection theory, so their data were excluded from analyses.

Results

Instructor evaluation ratings. Results closely paralleled those of Experiment 1. The fluent instructor was rated significantly higher than the disfluent instructor on organization, $t(100) = 4.61$, $p < .001$, $d = .91$, knowledge, $t(100) = 3.80$, $p < .001$, $d = .75$, preparedness, $t(100) = 5.61$, $p < .001$, $d = 1.10$, and overall effectiveness, $t(100) = 2.71$, $p = .008$, $d = .53$ (see Table 1). No significant differences were observed between the fluent and disfluent conditions in students' ratings of motivation (1.94 vs. 2.18, respectively), interest (1.58 vs. 1.92, respectively), or in how much they felt they learned (2.30 vs. 2.22, respectively), ts < 1.9.

Predicted versus actual performance. Test scores again revealed no significant difference in learning between the fluent condition ($M = 56\%$, $SD = 17\%$) and the disfluent condition ($M = 62\%$, $SD = 20\%$), $t(100) = 1.46$, $p = .15$. Response times associated with correct responses did not differ between the fluent condition ($M = 10.33$ s, $SD = 3.11$ s) and the disfluent condition ($M = 10.01$ s, $SD = 2.56$ s), $t(100) = .59$, $p = .56$. Judgments of learning were similar for the fluent condition ($M = 44\%$, $SD = 21\%$) and the disfluent condition ($M = 48\%$, $SD = 24\%$), $t(100) = .99$, $p = .33$. Thus, as in Experiment 1, students' confidence in their learning was not significantly affected by instructor fluency. The correlation between students' JOLs and test scores was again positive in both the fluent condition, $r = .32$, $p = .021$ and in the disfluent condition, $r = .33$, $p = .02$, indicating fairly consistent agreement between students' perceived learning and actual learning.

Discussion

Results of Experiments 1 and 2 reveal that a fluent instructor—accompanied by higher student ratings on traditional instructor evaluation questions such as organization and preparedness—does not appear to produce better learning than a disfluent instructor. This finding is consistent with previous research on instructor fluency using shorter videos of only 1 minute in length (Carpenter et al., 2013). Inconsistent with prior work, however, was the finding that students' judgments of learning did not differ between the two conditions. Students who viewed the fluent instructor predicted that they would score similarly on the upcoming test, on average, to those who viewed the disfluent instructor. Why might this be?

The answer may lie in the factor(s) that influence participants' judgments of learning. Unlike previous work by Carpenter et al. (2013), in which the videos were short and the instructor cues were salient, the current videos contained a higher degree of complexity that did not showcase the instructor as much. In the current videos, the instructor could only be heard and not seen, and her explanations were accompanied by fairly complex visual graphics and animations that helped illustrate the concepts. In comparison to the simplified videos used in previous work, it is likely that the current videos provided a greater variety of cues upon which participants could base their judgments.

The lack of a difference in judgments of learning between the two conditions could reflect the possibility that participants based their judgments on something other than the instructor; for example, the material itself and how difficult they perceived it to be. If

judgments of learning are based on nonmanipulated factors, systematic differences in judgments between the two conditions might not be expected to occur. On the other hand, if participants do base their judgments on the instructor—the factor that was manipulated—do differences emerge in students' judgments of learning between the fluent and disfluent conditions? If so, how do these differences coincide with actual learning?

Experiment 3

Experiment 3 was designed to answer these questions. The overall design was identical to the previous two experiments, except that after making their judgments of learning, participants were queried as to what factor(s) they believed formed the basis for their judgment. This permitted us to explore the frequency with which participants based their judgments on the instructor versus other, non-instructor-related factors and examine the results accordingly. The only other change to previous procedures was that the test in Experiment 3 was delayed by 24 hours, which represents a time interval between learning and test that is likely to occur in educational settings.

Participants

One hundred and six participants were recruited from the same participant pool as before. None of them had participated in Experiments 1 or 2.

Design and Procedure

As in Experiments 1 and 2, participants were randomly assigned to view either the fluent instructor video or the disfluent instructor video. They then made a judgment of learning (JOL) concerning how well they felt they would score (from 0% to 100%) on a multiple-choice test given on the information 24 hours later. Unlike in the previous experiments, immediately after making their JOL, participants were presented with the following instructions: "Think about the decision that you just made. On the following screen, we are going to ask you some questions about what formed the basis for your decision. Press the spacebar to begin." Participants were then presented with the following statements, one at a time (in randomized order for each participant): (a) "I based my decision on the material itself, and how difficult or easy I felt it would be to remember," (b) "I based my decision on the instructor who explained the material, and how good of a job I felt she did," (c) "I based my decision on my own general ability to learn and retain information," and (d) "I based my decision on something unrelated to the video, such as how sleepy or distracted I felt." On the screen below each statement, a scale from 1–6 appeared, where 1 indicated *strongly disagree*, and 6 indicated *strongly agree*. Participants indicated their agreement with each statement by pressing a number between 1 and 6. After indicating their agreement with all four statements, participants were given an open-ended question asking, "Is there anything else that you feel influenced your decision of how well you will score on tomorrow's test? If so, please give a brief description in the box below."

Immediately after making their JOLs and answering the questions about the bases for their JOLs, participants completed the same instructor evaluation questions from the previous experiments, and were then dismissed and reminded to return the next day for the test session. Upon returning for the test session, participants were given the same 20-item multiple-choice test from the previous experiments. Participants then answered a question about their prior knowledge of signal detection theory. Because the test session occurred on a different day from the learning session, participants were also asked if they had looked up or rehearsed any of the information since the learning session the previous day. After answering these questions, participants were thanked and debriefed.

Eight participants completed the learning session but failed to return for the test session. In addition, four participants indicated prior knowledge of signal detection theory, two indicated that they had looked up or rehearsed the material in between the learning session and the test session, and one participant failed to follow instructions during the learning phase. Data from these participants were excluded from all analyses, leaving 44 participants in the fluent condition and 47 participants in the disfluent condition.

Results and Discussion

Instructor evaluation ratings. Consistent with results from Experiments 1 and 2, the fluent instructor was rated significantly higher than the disfluent instructor on organization, $t(89) = 5.67$, $p < .001$, $d = 1.19$, knowledge, $t(89) = 2.82$, $p = .006$, $d = .61$, preparedness, $t(89) = 5.14$, $p < .001$, $d = 1.09$, and overall effectiveness, $t(89) = 5.46$, $p < .001$, $d = 1.14$ (see Table 1 for means and standard deviations). Students in the fluent condition also indicated higher ratings for how much they felt they had learned ($M = 3.32$, $SD = .74$) compared to students in the disfluent condition ($M = 2.77$, $SD = .98$), $t(89) = 3.01$, $p = .003$, $d = .63$. Also, a marginally significant difference emerged for motivation, with students in the fluent condition reporting higher motivation ($M = 2.95$, $SD = .94$) than students in the disfluent condition ($M = 2.55$, $SD = 1.02$), $t(89) = 1.95$, $p = .05$, $d = .41$. Ratings of interest in the material were not significantly different for participants in the fluent condition ($M = 2.55$, $SD = .82$) compared to the disfluent condition ($M = 2.23$, $SD = 1.00$), $t(89) = 1.61$, $p = .11$.

Predicted versus actual performance. Participants' JOLs were numerically higher in the fluent condition ($M = 63\%$, $SD = 16\%$) than in the disfluent condition ($M = 57\%$, $SD = 20\%$), but this difference was not significant, $t(89) = 1.55$, $p = .12$. Test scores, however, showed a small but reliable advantage for the fluent condition ($M = 60\%$, $SD = 20\%$) over the disfluent condition ($M = 52\%$, $SD = 17\%$), $t(89) = 2.07$, $p = .041$, $d = .43$. Response times associated with correct responses did not differ between the fluent condition ($M = 11.88$ s, $SD = 2.74$ s) and the disfluent condition ($M = 11.64$ s, $SD = 2.87$ s), $t(89) = .42$, $p = .68$. Thus, as in the previous experiments, participants' confidence in their learning was not significantly affected by instructor fluency. As before, the correlation between students' JOLs and test scores was positive in both the fluent condition, $r = .30$, $p = .046$ and in the disfluent condition, $r = .44$, $p = .002$, indicating fairly consistent agreement between students' perceived learning and actual learning.

The results of all three experiments indicate that participants do not appear to overestimate their own learning after watching a video of a fluent instructor versus a disfluent instructor, and their

JOLs in both conditions correlate positively with their later test scores. We hypothesized that the lack of difference in students' perceived learning between the two conditions may be due to the possibility that the instructor alone is not the primary cue upon which participants base their JOLs. If participants base their JOLs primarily on factors unrelated to the instructor—such as the material to be learned—this could explain why JOLs, on average, were not different between the two conditions. To explore this, we examined participants' responses to the questions regarding the factors that influenced their JOLs.

Factors influencing judgments of learning. Table 2 shows the proportion of participants who indicated 1 (*strongly disagree*) through 6 (*strongly agree*) in response to each of the factors that were queried. It appears that many participants did not endorse instructor as a strong basis for their judgments. Instead, participants often endorsed the material itself and their own general ability to learn and retain information. Across both conditions, more than 50% of participants gave a high agreement rating of 5 or 6 to these two factors. Ratings of 5 or 6 were only given in response to the instructor as the basis for the judgments by 27% of participants in the fluent condition, and by 45% of participants in the disfluent condition.[2]

When participants do base their JOLs on the instructor, are they more likely to exhibit overconfidence after viewing a fluent instructor compared to a disfluent instructor? To answer this question, we examined the data only for those participants who endorsed instructor (i.e., gave a rating of 5 or 6) as a basis for their judgments in the fluent condition ($n = 12$) and in the disfluent condition ($n = 21$). Figure 2 displays the mean predicted test scores (i.e., JOLs) and actual test scores across the two conditions for these 33 participants. A 2 × 2 (Performance: Predicted vs. Actual × Condition: Fluent vs. Disfluent) Mixed ANOVA revealed a significant interaction, $F(1, 31) = 4.31$, $p = .046$, $\eta^2 = .12$, indicating that predicted performance exceeded actual performance more so for participants in the fluent condition ($t = 2.99$, $p = .012$, $d = .86$) than in the disfluent condition, $t = .47$, $p = .64$. This interaction—the same one reported by Carpenter et al. (2013)—indicates that instructor fluency, when used as a basis for judgments of learning, can lead to inflated estimates of one's own learning.

Table 2
Proportion of Participants Responding 1–6 to Factors That Influenced Their Judgments of Learning in Experiment 3

	Ratings (1 = *strongly disagree*, 6 = *strongly agree*)					
	1	2	3	4	5	6
Fluent instructor						
Instructor	.09	.23	.16	.25	.25	.02
Material	.02	.05	.11	.30	.25	.27
Ability	.00	.07	.14	.27	.43	.09
Other	.02	.16	.16	.36	.18	.12
Disfluent instructor						
Instructor	.06	.15	.06	.28	.28	.17
Material	.00	.02	.28	.19	.28	.23
Ability	.02	.07	.19	.17	.32	.23
Other	.04	.21	.09	.23	.32	.11

Note. After estimating how well they believed they would score on the upcoming test (from 0 to 100%), participants were asked to rate their agreement (from 1 to 6) with each of the following statements: (a) "I based my decision on the instructor who explained the material, and how good of a job I felt she did," (b) "I based my decision on the material itself, and how difficult or easy I felt it would be to remember," (c) "I based my decision on my own general ability to learn and retain information," and (d) "I based my decision on something unrelated to the video, such as how sleepy or distracted I felt."

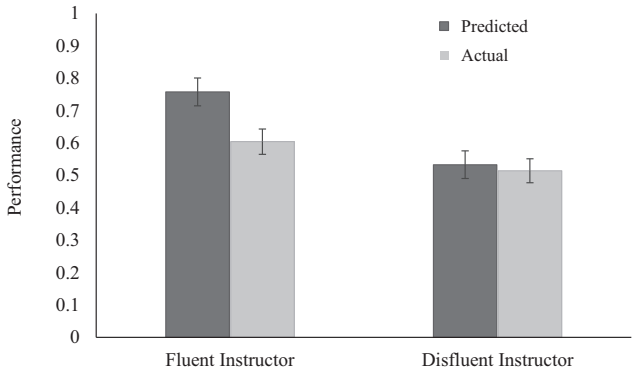

Figure 2. Predicted performance versus actual performance for participants who endorsed instructor (gave a rating of 5 or 6) as a basis for their judgments of learning in Experiment 3.

We examined the same effect by performing a continuous analysis of the data that included all participants. For each participant, a calibration score was computed by subtracting actual test performance from predicted test performance. The resulting value indicates the degree to which each participant was overconfident (where the predicted score is higher than the actual score, reflected by a positive value) or underconfident (where the predicted score is lower than the actual score, reflected by a negative value). These calibration scores ranged from .65 (one participant who predicted a test score of 75%, but only scored 10% on the test) to −.30 (one participant who predicted a test score of 30%, but scored 60% on the test).

Each participant's calibration score was correlated with the rating that they gave (1–6) indicating the degree to which they based their JOL on the instructor. In the fluent condition, a Spearman rank order correlation coefficient between these two measures indicated that greater reliance on the instructor as the basis for JOLs coincided with greater overconfidence, $r_s(44) = .34$, $p = .026$. In the disfluent condition, the same correlation was negative (but nonsignificant), indicating that greater reliance on the instructor as the basis for JOLs coincided with underconfidence, $r_s(47) = -.17$, $p = .26$. Consistent with the interaction reported above, these correlations indicate that the more participants rely on the instructor as a basis for their JOLs, the more likely they are to be overconfident when learning from a fluent instructor.

The same Spearman correlations revealed no significant relationships in either condition between calibration scores and

[2] Participants in the disfluent condition based their JOLs on the instructor more often than did participants in the fluent condition. The most likely reason for this is that the disfluent condition contained fairly noticeable vocal cues of disfluency (e.g., stammering, pauses, and frequent use of "ums") that were not present in the fluent condition. These vocal cues likely drew more attention to the instructor in the disfluent condition than in the fluent condition.

the degree to which participants relied on the material itself as a basis for their JOLs ($r_{s\ fluent} = -.12$, $r_{s\ disfluent} = -.05$), or on their own ability to learn and retain information as a basis for their JOLs ($r_{s\ fluent} = .17$, $r_{s\ disfluent} = .13$), all ps $> .27$. The degree to which participants based their JOLs on something unrelated to the video (e.g., how sleepy or distracted they felt) coincided with underconfidence in both the fluent condition ($r_s = -.30$, $p = .048$) and in the disfluent condition ($r_s = -.13$, $p = .40$). This most likely reflected participants' deflated sense of confidence in their ability to retain information under conditions where they felt their learning was influenced by external factors.[3]

Thus, when participants base their JOLs on the instructor, they are more likely to be overconfident in their own learning after viewing a fluent instructor compared to a disfluent instructor. When they base their JOLs on the material itself, or on their own general ability to learn and retain information, they exhibited no significant biases—neither overconfidence nor underconfidence. These results suggest that basing one's judgments on the material or on one's own abilities may prevent systematic errors in assessing one's own learning, but basing one's judgments on a fluent instructor can lead to overconfidence.

General Discussion

The current study adds new data to our understanding of the influence of instructor fluency on students' perceptions of instructors, confidence in their own learning, and their actual learning. Across three experiments, we found that a fluent instructor was rated significantly higher on traditional instructor evaluation questions measuring organization, preparedness, knowledge, and overall effectiveness. This is consistent with previous work showing that the behavior of an instructor—even if it is unrelated to the content being learned—can significantly influence students' perceptions of instructors (Carpenter et al., 2013).

A similar effect has been observed in studies manipulating instructor expressiveness. When an instructor delivers a lecture that contains gestures, humor, and personal anecdotes, students' evaluations of instructors are higher than when the same lecture topic is delivered by the same instructor without these expressive behaviors (e.g., Ware & Williams, 1975; Williams & Ware, 1976, 1977). Although the presence of jokes and personal stories in one condition and not the other means that the material being presented was not always identical in these studies, the extra information in the "expressive" condition was unrelated to the content being taught, meaning that students' perceptions of instructors can be based on factors that have nothing to do with what they are learning about.

An extreme example of this—thereafter referred to as the "Dr. Fox Effect" (Ware & Williams, 1975)—was demonstrated by Naftulin, Ware, and Donnelly (1973). In this study, researchers arranged a live guest lecture to be given to an audience of medical educators during a teacher training conference. The topic was on mathematical game theory applied to medical education, and the speaker was Dr. Myron L. Fox, who was introduced as an expert on mathematics and human behavior. Unbeknownst to the audience, "Dr. Fox" was really a Hollywood actor who knew nothing about game theory or medical education. He prepared the lecture from a brief, 5-page article in Scientific American geared toward lay readers (Rapoport, 1962), and he was instructed to present the topic in a way that the content itself would be meaningless. This was accomplished by including references to vague and abstract things that were never clarified, frequent use of humorous stories unrelated to the topic, redundant points, and multiple contradictory statements. Dr. Fox delivered the meaningless 1-hr lecture in a way that conveyed a sense of authority on the topic and a high degree of enthusiasm. Afterward, an evaluation questionnaire filled out by the audience indicated overwhelmingly positive impressions of the lecture. Over 90% of audience members felt that it was interesting and well-organized, contained good examples to clarify the material, and stimulated them to think more about the topic. In their open-ended statements, audience members made no mention of the vague material or the contradictory statements, and after being informed about the study, none of them reported ever suspecting that the lecture was a hoax.

Findings like these highlight the important distinction between students' impressions of instructors and their learning of meaningful content. Even with a much more subtle manipulation of instructor behavior, the current study confirms that students are sensitive to behaviors of the instructor that reflect a sense of preparedness, organization, and knowledge of the topic. However, our results indicate that students may be aware that a positive impression of an instructor, as reflected by these factors, does not necessarily mean better learning. Though fairly strong differences were observed in ratings of instructor effectiveness between the fluent and disfluent conditions, judgments of learning on average were no different across the two conditions. In Experiment 3, when queried about the factors that influenced their judgments, students in both conditions most often reported that their judgments were influenced by the material itself and their own general ability to learn and retain information. Thus, even if students feel that an instructor is very knowledgeable, engaging, and has all the qualities of a "good" instructor, they can dissociate their perceptions of the instructor from how much they feel they have learned. This appears to be especially true if they base their judgments of learning on the material that they are learning, or on their own perceived abilities to learn and retain information.

Students' tendency to base their judgments of learning on the material itself could explain the finding that their judgments, on average, coincided fairly well with their actual test scores. Previ-

[3] We also examined participants' open-ended responses as to what additional factors, if any, influenced their JOLs. This question received a response by 50% of participants in the fluent condition and 34% of participants in the disfluent condition. Examination of these responses revealed that they were largely consistent with the options that participants had already responded to. For example, one participant typed "The things that influenced me are how tired I am and how I feel I do not retain information easily in general." Consistent with this description, the same participant gave high ratings to the *ability* (5) and *unrelated* (6) factors, and lower ratings to the *instructor* (1) and *material* (4) factors. Each open-ended response across the two conditions was coded, in blind fashion, as to whether it described the instructor ($n = 13$ responses), the material ($n = 21$ responses), participants' abilities ($n = 5$ responses), or something unrelated to the video ($n = 5$ responses). Across the two conditions, responses that were coded as fitting into each of these categories coincided with a median response rating of 5 for each category, reflecting high agreement between participants' open-ended responses and the ratings they gave when the options were presented to them. Responses that reflected "unrelated to the video" included feeling tired, thinking about homework or other classes, having just come from a long lecture class, and not being able to take notes during the lecture video.

ous research has shown that students' judgments of learning can be sensitive to the difficulty of the material being learned, which can in turn directly influence performance. When material is made objectively more difficult—for example, by preselecting trivia questions to be of high versus low difficulty (Pulford & Colman, 1997), or by altering the coherency of a text passage to make it harder to read (Rawson & Dunlosky, 2002)—students express lower confidence in their ability to remember the information, and do indeed remember it less well on a subsequent test. In the current study, the concepts associated with signal detection theory may have been perceived as difficult, leading students to express lower confidence overall than what would be expected based on previous research using simpler types of stimuli such as word lists (Alter et al., 2007; Castel et al., 2007; Kornell & Bjork, 2009; Rhodes & Castel, 2008) or familiar pictures (Carpenter & Olson, 2012), which usually induce overconfidence.

Thus, students may not automatically fall prey to overconfidence when learning from fluent instructors. If the content itself is somewhat challenging, students may use their perception of the content as a primary cue in assessing their own knowledge, and this may be a better indication of how they will perform on a future test than is the behavior of the instructor. We did find, however, that a portion of students relied upon the instructor as a basis for their judgment, indicating that challenging content alone does not inoculate students from potentially misleading metacognitive cues. When students based their judgments of learning on the instructor in Experiment 3, they were overconfident in their judgments to a greater degree after viewing the fluent instructor compared to the disfluent instructor. This result is consistent with that observed by Carpenter et al. (2013), and indicates that instructor fluency, when relied upon as a basis for one's judgment of learning, can induce an illusion of knowing.

To reduce these illusions and encourage students' reliance on cues that are more diagnostic of their actual learning, instructors may find it useful to incorporate into lectures techniques that are known to improve students' metacognitive monitoring. One such technique is retrieval practice, which has been shown to improve the accuracy of students' predictions about their own performance (e.g., Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Carpenter et al., 2015; Finn & Metcalfe, 2007; Little & McDaniel, 2015; Szpunar, Jing, & Schacter, 2014; Tauber & Rhodes, 2012; Tullis, Finley, & Benjamin, 2013). This technique is useful in general for helping students learn (e.g., Butler, Marsh, Slavinsky, & Baraniuk, 2014; Carpenter, 2012; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Rowland, 2014; Roediger & Butler, 2011), and it may introduce the added benefit of dispelling illusions of knowing that could arise when students view a smooth and well-polished lecture that may, on the surface, look easy to learn.

Does the fluency of an instructor reliably affect students' actual learning? We found that students' test scores did not differ consistently following a fluent versus disfluent lecture. In Experiment 1 test scores were similar between the two conditions (66% vs. 65%), in Experiment 2 they were numerically (but not significantly) higher following the disfluent lecture (62%) than the fluent lecture (56%), and in Experiment 3 there was a small but significant advantage for the fluent lecture (60%) over the disfluent lecture (52%). We hypothesized that with educationally relevant materials, fluent lectures might lead to better learning than disfluent lectures due to the boredom or disengagement that would seem more likely to occur during a disfluent lecture. Though the data from Experiment 3 are suggestive of this possibility, the difference was small and did not occur in the other two experiments. Learning decrements associated with disfluent lectures, therefore, do not appear to be particularly pervasive under these conditions, though it is possible that factors yet to be systematically explored (e.g., the length of the delay between learning and testing, or the nature of the materials to be learned) could reveal such decrements.

Paralleling the current results are the findings from several earlier studies on instructor expressiveness. Driven in large part by the findings of the "Dr. Fox study" (Naftulin et al., 1973), these studies compared student learning from lectures that were delivered by the same instructor in a style that was high in expressiveness (use of gestures, humor, and personal anecdotes) versus low in expressiveness (minimizing or eliminating these things). Students then rated the instructor using traditional measures of instructor evaluation—the degree to which the instructor displayed enthusiasm, presented the material clearly, and was well prepared for the lecture—and then took a test on the content that was taught in the lecture. These studies found that instructors who demonstrated high expressiveness were rated higher by students, but were not associated with significantly higher test scores than instructors who demonstrated low expressiveness (Meier & Feldhusen, 1979; Perry, Abrami, & Leventhal, 1979; Williams & Ware, 1976).

Exceptions were reported by Ware and Williams (1975) and Williams and Ware (1977), who found that students did score significantly higher (by about 10%) on a quiz following the high-expressive instructor compared to the low-expressive instructor. Coats and Smidchens (1966) also reported that students' immediate recall of information was significantly higher following a dynamic lecture (i.e., the speaker using gestures and vocal inflections, moving around, and presenting the information from memory) versus a static lecture given by the same person (i.e., reading the lecture from notes, minimizing eye contact and vocal inflections). The reasons for these different findings are presently not clear. Many of the studies on this topic manipulated a number of additional variables beyond instructor expressiveness, including coverage of the content in the lecture (Meier & Feldhusen, 1979; Ware & Williams, 1975; Williams & Ware, 1976, 1977), incentives for students to learn (Coats & Smidchens, 1966; Perry et al., 1979; Williams & Ware, 1976), and whether or not students had additional opportunities to study the content after viewing the lecture and taking the quiz (Perry et al., 1979). No consistent interactions emerged from these manipulations to identify the conditions under which an instructor's degree of expressiveness might benefit learning. However, the results of a related study by Leventhal, Perry, and Abrami (1977) indicated that students' quiz scores were higher following a lecture given by an enthusiastic instructor who made frequent use of the blackboard to explain concepts, versus the same instructor who delivered the lecture without displaying these behaviors. This advantage, however, only occurred when students were led to believe that the instructor was inexperienced. For students who were told that they were viewing an experienced instructor who had been teaching for many years, quiz scores were no different whether the instructor was enthusiastic and dynamic, versus static and subdued. These results raise the interesting possibility that the effect of an instructor's behavior on student learning may depend on particular student characteristics, such as preexisting beliefs and expectations.

Thus, the current state of research suggests that instructor behaviors based on fluency or expressiveness do not appear to have a strong and consistent effect on learning. Given the somewhat mixed results, along with the fact that few studies have been conducted on this topic, an exciting and worthwhile endeavor for future research is to further explore the effects of instructor fluency, particularly geared toward identifying moderating factors that may determine the conditions under which instructor fluency benefits learning. It may be worthwhile to explore whether such effects are influenced by the timing of students' judgments of learning (e.g., whether judgments are made immediately after learning, or sometime later such as just prior to taking a test), and the level of complexity of the knowledge that students are tested on. Future research may also find it worthwhile to explore the indirect effects of instructor fluency. The research reported here was concerned with the direct effects of instructor fluency. If instructor fluency does not affect learning in direct ways, it seems quite possible that it may affect learning in indirect ways, perhaps through increased absences or a lack of interest in the material that leads to less studying. It is also possible that in authentic learning situations, the fluency of an instructor's style is related to the quality of the content presented—instructors who are well-prepared and organized may have higher-quality content than those who are less prepared and less organized—and the combination of content and delivery style could affect students' learning. Future research that can shed additional light on this interesting and empirically wide-open topic is highly encouraged.

One thing that is clear from this research is that instructor fluency has a greater effect on students' ratings of instructors than it does on students' learning. This result carries important implications for students' evaluations of instructors. Student evaluations have long been used as a means of measuring the quality of teaching in colleges and universities. Based on the reasonable assumption that students—as the recipients of instruction—are in the best position to evaluate its effectiveness, input is collected year after year concerning students' perceptions of their courses and the instructors who have taught them. Past and current findings converge, however, to suggest that student evaluations of instructors may not be the most accurate indicator of how much students have learned.

Data collected from actual courses appear to corroborate this. While some data suggest that students' evaluations of instructors are positively correlated with the grades they receive in the courses taught by those instructors (Marsh, Fleiner, & Thomas, 1975), these data cannot rule out the possibility that students' perceptions of instructors were influenced by the grade they were receiving at the time the rating was made. To avoid this problem, students' knowledge of content from a particular course has sometimes been tested using a standardized assessment that was prepared and administered by someone other than the instructor of that course. Using this method, some studies have shown a positive relationship between instructor ratings and knowledge gained from the course (Bryson, 1974; Sullivan & Skanes, 1974), some studies have shown no relationship (Galbraith, Merrill, & Kline, 2012; Palmer, Carliner, & Romer, 1978), and some studies have even shown a negative relationship (Yunker & Yunker, 2003). These findings demonstrate that there are many factors that could influence students' learning—including the content of the course, difficulty of the material, and size of the class—and these factors may or may not coincide with the perceived effectiveness of the instructor who taught them. Further studies have shown that students' ratings of instructors can be based on a variety of non-course-related factors such as the personality, gender, age, and attractiveness of the instructor (Abrami, Leventhal, & Perry, 1982; Goebel & Cashen, 1979; Neath, 1996).

Thus, student ratings should be interpreted with caution if used as a means of assessing whether instructors are enhancing students' learning. Recent survey data indicate that 87% of university administrators report using student ratings to inform personnel decisions about instructors, such as promotion, tenure, and merit pay (Beran, Violato, Kline, & Frideres, 2005). If relied upon as the primary measure of an instructor's effectiveness, such ratings could give a biased impression that might influence these important decisions. As such, some researchers have advocated for the use of additional sources of data—such as objective measures of student achievement or peer evaluations—that can supplement the information gained from student ratings (Emery, Kramer, & Tian, 2003). Other suggestions for optimizing the use of student ratings and other data to measure teaching effectiveness have recently been discussed by Gravestock and Gregor-Greenleaf (2008) and Wright (2006).

In closing, we note that although students' perceptions of instructors do not appear to consistently coincide with learning, these perceptions can still provide valuable information on other aspects of teaching that are useful to students, instructors, and administrators. For example, students' input can reveal potential accountability issues such as an instructor's persistent tardiness or failure to fulfill responsibilities, and can likewise reveal positive examples such as the acknowledgment of outstanding mentors. A positive perception of an instructor may also inspire students to take more classes in a given area or choose a particular career path. Instructors who are highly regarded by students may influence those students in a number of ways that are not restricted to the learning of particular subject matter from a course. The relationship between students' perceptions of instructors and their educational experiences—both objective and subjective—is likely to be a multifaceted one that is currently not well understood and deserving of further research.

References

Abrami, P. C., Leventhal, L., & Perry, R. P. (1982). Educational seduction. *Review of Educational Research, 52,* 446–464. http://dx.doi.org/10.3102/00346543052003446

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., III, & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22,* 861–876. http://dx.doi.org/10.1002/acp.1391

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review, 13,* 219–235. http://dx.doi.org/10.1177/1088868309341564

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General, 136,* 569–576. http://dx.doi.org/10.1037/0096-3445.136.4.569

Baillie, C., & Fitzgerald, G. (2000). Motivation and attrition in engineering students. *European Journal of Engineering Education, 25,* 145–155. http://dx.doi.org/10.1080/030437900308544

Beran, T., Violato, C., Kline, D., & Frideres, J. (2005). The utility of student ratings of instruction for students, faculty, and administrators: A

"consequential validity" study. *Canadian Journal of Higher Education, 35,* 49–70.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64,* 417–444. http://dx.doi.org/10.1146/annurev-psych-113011-143823

Bol, L., Hacker, J. D., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *Journal of Experimental Education, 73,* 269–290. http://dx.doi.org/10.3200/JEXE.73.4.269-290

Brown, G., & Atkins, M. (1990). *Effective teaching in higher education.* London, UK: Routledge.

Brown, S., & Race, P. (2002). *Lecturing: A practical guide*. Sterling, VA: Stylus Publishing Inc.

Bryson, R. (1974). Teacher evaluations and student learning: A reexamination. *The Journal of Educational Research, 68,* 12–14. http://dx.doi.org/10.1080/00220671.1974.10884692

Butler, A. C., Marsh, E. J., Slavinsky, J. P., & Baraniuk, R. G. (2014). Integrating cognitive science and technology improves learning in a STEM classroom. *Educational Psychology Review, 26,* 331–340. http://dx.doi.org/10.1007/s10648-014-9256-4

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21,* 279–283. http://dx.doi.org/10.1177/0963721412452728

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H. K., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review, 24,* 369–378. http://dx.doi.org/10.1007/s10648-012-9205-z

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name-learning. *Applied Cognitive Psychology, 19,* 619–636. http://dx.doi.org/10.1002/acp.1101

Carpenter, S. K., Lund, T. J. S., Coffman, C. R., Armstrong, P. I., Lamm, M. H., & Reason, R. D. (2015). A classroom study on the relationship between student achievement and retrieval-enhanced learning. *Educational Psychology Review*. Advance online publication. http://dx.doi.org/10.1007/s10648-015-9311-9

Carpenter, S. K., & Mueller, F. E. (2013). The effects of interleaving versus blocking on foreign language pronunciation learning. *Memory & Cognition, 41,* 671–682. http://dx.doi.org/10.3758/s13421-012-0291-4

Carpenter, S. K., & Olson, K. M. (2012). Are pictures good for learning new vocabulary in a foreign language? Only if you think they are not. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 92–101. http://dx.doi.org/10.1037/a0024828

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using testing to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology, 23,* 760–771. http://dx.doi.org/10.1002/acp.1507

Carpenter, S. K., Wilford, M. M., Kornell, N., & Mullaney, K. M. (2013). Appearances can be deceiving: Instructor fluency increases perceptions of learning without increasing actual learning. *Psychonomic Bulletin & Review, 20,* 1350–1356. http://dx.doi.org/10.3758/s13423-013-0442-z

Castel, A. D., McCabe, D. P., & Roediger, H. L., III. (2007). Illusions of competence and overestimation of associative memory for identical items: Evidence from judgments of learning. *Psychonomic Bulletin & Review, 14,* 107–111. http://dx.doi.org/10.3758/BF03194036

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science, 19,* 1095–1102. http://dx.doi.org/10.1111/j.1467-9280.2008.02209.x

Coats, W. D., & Smidchens, U. (1966). Audience recall as a function of speaker dynamism. *Journal of Educational Psychology, 57,* 189–191. http://dx.doi.org/10.1037/h0023568

Davis, B. G. (1993). *Tools for teaching*. San Francisco, CA: Wiley.

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. *Psychology of Learning and Motivation, 53,* 63–147. http://dx.doi.org/10.1016/S0079-7421(10)53003-2

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Beverly Hills, CA: Sage.

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory and Language, 33,* 545–565. http://dx.doi.org/10.1006/jmla.1994.1026

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14,* 4–58. http://dx.doi.org/10.1177/1529100612453266

Ekeler, W. J. (1994). The lecture method. In K. W. Prichard & R. M. Sawyer (Eds.), *Handbook of college teaching: Theory and applications* (pp. 85–98). Westport, CT: Greenwood.

Emery, C. R., Kramer, T. R., & Tian, R. G. (2003). Return to academic standards: A critique of student evaluations of teaching effectiveness. *Quality Assurance in Education, 11,* 37–46. http://dx.doi.org/10.1108/09684880310462074

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 238–244. http://dx.doi.org/10.1037/0278-7393.33.1.238

Finn, B., & Tauber, S. K. (2015). When confidence is not a signal of knowing: How students' experiences and beliefs about processing fluency can lead to miscalibrated confidence. *Educational Psychology Review, 27,* 567–586. http://dx.doi.org/10.1007/s10648-015-9313-7

Galbraith, C. S., Merrill, G. B., & Kline, D. M. (2012). Are student evaluations of teaching effectiveness valid for measuring student learning outcomes in business related classes? A neural network and Bayesian analysis. *Research in Higher Education, 53,* 353–374. http://dx.doi.org/10.1007/s11162-011-9229-0

Geisinger, B. N., & Raman, D. R. (2013). Why they leave: Understanding student attrition from engineering majors. *International Journal of Engineering Education, 29,* 914–925.

Gerbier, E., & Toppino, T. C. (2015). The effect of distributed practice: Neuroscience, cognition, and education. *Trends in Neuroscience & Education, 4,* 49–59. http://dx.doi.org/10.1016/j.tine.2015.01.001

Goebel, B. L., & Cashen, V. M. (1979). Age, sex, and attractiveness as factors in student ratings of teachers: A developmental study. *Journal of Educational Psychology, 71,* 646–653. http://dx.doi.org/10.1037/0022-0663.71.5.646

Gravestock, P., & Gregor-Greenleaf, E. (2008). *Student course evaluations: Research, models, and trends*. Toronto, ON: Higher Education Quality Council of Ontario.

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92,* 160–170. http://dx.doi.org/10.1037/0022-0663.92.1.160

Hogan, J. (1999). Lecturing for learning. In H. Fry, S. Ketteridge, & S. Marshall (Eds.), *A handbook for teaching and learning in higher education: Enhancing academic performance* (pp. 83–94). New York, NY: Routledge.

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 187–194. http://dx.doi.org/10.1037/0278-7393.31.2.187

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131,* 147–162.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14,* 219–224. http://dx.doi.org/10.3758/BF03194055

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science, 19,* 585–592. http://dx.doi.org/10.1111/j.1467-9280.2008.02127.x

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General, 138,* 449–468. http://dx.doi.org/10.1037/a0017350

Küpper-Tetzel, C. E. (2014). Understanding the distributed practice effect: Strong effects on weak theoretical grounds. *Zeitschrift fur Psychologie, 222,* 71–81. http://dx.doi.org/10.1027/2151-2604/a000168

Leventhal, L., Perry, R. P., & Abrami, P. C. (1977). Effects of lecturer quality and student perception of lecturer's experience on teacher ratings and student achievement. *Journal of Educational Psychology, 69,* 360–374. http://dx.doi.org/10.1037/0022-0663.69.4.360

Little, J. L., & McDaniel, M. A. (2015). Metamemory monitoring and control following retrieval practice for text. *Memory & Cognition, 43,* 85–98. http://dx.doi.org/10.3758/s13421-014-0453-7

Lowman, J. (1995). *Mastering the techniques of teaching* (2nd ed.). San Francisco, CA: Jossey-Bass.

Marsh, H. W., Fleiner, H., & Thomas, C. S. (1975). Validity and usefulness of student evaluations of instructional quality. *Journal of Educational Psychology, 67,* 833–839. http://dx.doi.org/10.1037/0022-0663.67.6.833

Meier, R. S., & Feldhusen, J. F. (1979). Another look at Dr. Fox: Effect of stated purpose for evaluation, lecture expressiveness, and density of lecture content on student ratings. *Journal of Educational Psychology, 71,* 339–345. http://dx.doi.org/10.1037/0022-0663.71.3.339

Miller, T. M., & Geraci, L. (2011). Unskilled but aware: Reinterpreting overconfidence in low-performing students. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 502–506. http://dx.doi.org/10.1037/a0021802

Morton, A. (2009). Lecturing to large groups. In H. Fry, S. Ketteridge, & S. Marshall (Eds.), *A handbook for teaching and learning in higher education: Enhancing academic practice* (pp. 58–71). New York, NY: Routledge.

Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review, 20,* 378–384. http://dx.doi.org/10.3758/s13423-012-0343-6

Naftulin, D. H., Ware, J. E., Jr., & Donnelly, F. A. (1973). The Doctor Fox Lecture: A paradigm of educational seduction. *Journal of Medical Education, 48,* 630–635.

Neath, I. (1996). How to improve your teaching evaluations without improving your teaching. *Psychological Reports, 78,* 1363–1372. http://dx.doi.org/10.2466/pr0.1996.78.3c.1363

Palmer, J., Carliner, G., & Romer, T. (1978). Leniency, learning, and evaluations. *Journal of Educational Psychology, 70,* 855–863. http://dx.doi.org/10.1037/0022-0663.70.5.855

Perry, R. P., Abrami, P. C., & Leventhal, L. (1979). Educational seduction: The effect of instructor expressiveness and lecture content on student ratings and achievement. *Journal of Educational Psychology, 71,* 107–116. http://dx.doi.org/10.1037/0022-0663.71.1.107

Pulford, B. D., & Colman, A. M. (1997). Overconfidence: Feedback and item difficulty effects. *Personality and Individual Differences, 23,* 125–133. http://dx.doi.org/10.1016/S0191-8869(97)00028-7

Rapoport, A. (1962). The use and misuse of game theory. *Scientific American, 207,* 108–118. http://dx.doi.org/10.1038/scientificamerican1262-108

Rawson, K. A., & Dunlosky, J. (2002). Are performance predictions for text based on ease of processing? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 69–80. http://dx.doi.org/10.1037/0278-7393.28.1.69

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General, 137,* 615–625. http://dx.doi.org/10.1037/a0013684

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15,* 20–27. http://dx.doi.org/10.1016/j.tics.2010.09.003

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140,* 1432–1463. http://dx.doi.org/10.1037/a0037559

Serra, M. J., & Dunlosky, J. (2010). Metacomprehension judgements reflect the belief that diagrams improve learning from text. *Memory, 18,* 698–711. http://dx.doi.org/10.1080/09658211.2010.506441

Simon, D. A., & Bjork, R. A. (2001). Metacognition in motor learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 907–912.

Sullivan, A. M., & Skanes, G. R. (1974). Validity of student evaluation of teaching and the characteristics of successful instructors. *Journal of Educational Psychology, 66,* 584–590. http://dx.doi.org/10.1037/h0036929

Szpunar, K. K., Jing, H. G., & Schacter, D. L. (2014). Overcoming overconfidence in learning from video-recorded lectures: Implications of interpolated testing for online education. *Journal of Applied Research in Memory & Cognition, 3,* 161–164. http://dx.doi.org/10.1016/j.jarmac.2014.02.001

Tauber, S. K., & Rhodes, M. G. (2012). Multiple bases for young and older adults' judgments of learning in multitrial learning. *Psychology and Aging, 27,* 474–483. http://dx.doi.org/10.1037/a0025246

Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory & Cognition, 41,* 429–442. http://dx.doi.org/10.3758/s13421-012-0274-5

Ware, J. E., Jr., & Williams, R. G. (1975). The Dr. Fox effect: A study of lecturer effectiveness and ratings of instruction. *Journal of Medical Education, 50,* 149–156.

Williams, R. G., & Ware, J. E. (1976). Validity of students' ratings of instruction under different incentive conditions: A further study of the Dr. Fox effect. *Journal of Educational Psychology, 68,* 48–56. http://dx.doi.org/10.1037/0022-0663.68.1.48

Williams, R. G., & Ware, J. E. (1977). An extended visit with Dr. Fox: Validity of student satisfaction with instruction ratings after repeated exposures to a lecture. *American Educational Research Journal, 14,* 449–457. http://dx.doi.org/10.3102/00028312014004449

Wright, R. E. (2006). Student evaluations of faculty: Concerns raised in the literature, and possible solutions. *College Student Journal, 40,* 417–422.

Yunker, P. J., & Yunker, J. A. (2003). Are student evaluations of teaching valid? Evidence from an analytical business core course. *Journal of Education for Business, 78,* 313–317. http://dx.doi.org/10.1080/08832320309598619

Received September 26, 2015
Revision received December 10, 2015
Accepted December 13, 2015 ■