Journal of Applied Research in Memory and Cognition 3 (2014) 161–164

Journal of Applied Research in Memory and Cognition

Overcoming overconfidence in learning from video-recorded lectures: Implications of interpolated testing for online education

Karl K. Szpunar*, Helen G. Jing, Daniel L. Schacter

Harvard University, Department of Psychology, United States

ARTICLE INFO

Article history:
Received 3 December 2013
Received in revised form 20 February 2014
Accepted 25 February 2014
Available online 3 March 2014

Keywords:
Interpolated testing
Online learning
Video-recorded lectures
Judgments of learning

ABSTRACT

The video-recorded lecture represents a central feature of most online learning platforms. Nonetheless, little is known about how to best structure video-recorded lectures in order to optimize learning. Here, we focused on the tendency for high school and college students to be overconfident in their learning from video-recorded modules, and demonstrated that testing could be used to effectively improve the calibration between predicted and actual performance. Notably, interpolating a lecture with repeated tests helped to boost actual performance to the level of predicted performance, whereas a single test following the lecture served to lower unrealistic judgments of learning. The value of improving performance to match predictions of learning and other avenues for future research regarding meta-comprehension of video-recorded lectures is discussed.

1. Introduction

The video-recorded lecture represents a central feature of most online learning platforms (Breslow et al., 2013). Nonetheless, little remains known about what obstacles students might encounter when learning from video-recorded lectures or how those obstacles might be overcome. Here, we focus on how well students think they will perform on a later assessment associated with learning from a video-recorded lecture. Considerable research has indicated that students overestimate their ability to assess later performance associated with learning from video-recorded modules (Choi & Johnson, 2005; Salomon, 1984; for a recent review, see Means, Toyama, Murphy, Bakia, & Jones, 2010). Importantly, overconfidence in later performance can have a negative impact on long-term retention. For instance, students making overconfident judgments of learning have been shown to cut short subsequent opportunities for re-study (Dunlosky & Rawson, 2012; see also Bol & Hacker, 2012).

Further complicating matters, students tend to hold stable persistent beliefs about how well they learn in traditional educational settings (Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008; Schraw, Potenza, & Nebelsick-Gullet, 1993). For instance, various studies have demonstrated that it can be difficult to alter global judgments of learning that are based, at least partly, on lecture content (e.g., judgments of learning for mid-term exams; for a recent review, see Hacker, Bol, & Keener, 2008). Given that students overestimate how well they will perform on subsequent assessments associated with video-recorded materials and that this metacognitive error may be difficult to correct, what can be done to improve calibration between predicted and actual performance in online learning environments?

One approach may be to seek interventions that re-structure lectures in a manner that can boost actual performance to the level of predicted performance. Along these lines, considerable research has demonstrated that the act of retrieving information from memory can serve to boost learning and retention in educational settings (for relevant reviews, see Roediger & Butler, 2011; Roediger & Karpicke, 2006). Indeed, we recently demonstrated that interpolating a video-recorded lecture with brief memory tests served to substantially enhance learning (Szpunar, Khan, & Schacter, 2013). In the present study, we sought to examine whether interpolated testing during a lecture would elevate actual performance to the level of predicted performance.

The study involved three groups of high school students who learned from a statistics lecture. The use of video-recorded lectures in the context of online learning is quickly becoming a popular method of delivering educational content with high school populations (Picciano, Seaman, Shea, & Swan, 2012). Moreover, statistics is commonly perceived as being especially difficult to master (Gal & Ginsburg, 1994), and so any indication of overconfidence in learning from a statistics lecture would further highlight the robust

* Corresponding author at: Department of Psychology, Harvard University, 33 Kirkland Street, Cambridge, MA 02138, USA. Tel.: +1 617 495 9031.
E-mail address: szpunar@wjh.harvard.edu (K.K. Szpunar).

http://dx.doi.org/10.1016/j.jarmac.2014.02.001
2211-3681/

nature of overconfidence in learning from video-recorded materials.

One group of high school students learned the lecture in the presence of interpolated tests, whereas another group of high school students learned the lecture in the absence of interpolated tests. Although we did not expect global judgments of learning to differ between the two groups, we predicted that interpolated testing should nonetheless boost final test scores in a manner that would better align predicted and actual performance. Hence, students in the interpolated group should appear less overconfident and better calibrated. An alternative hypothesis is that interpolated testing could boost both actual and predicted performance and hence not improve calibration. Specifically, it is possible that students may find the act of answering questions during the lecture easy, which could serve to elevate predictions of future performance (cf. Schunk, 1991). To test the generalizability of the effects of interpolated testing on actual and predicted performance, we also carried out a partial re-analysis of an existing dataset that involved college students learning from the same statistics lecture under conditions of interpolated and non-interpolated testing.

Finally, we included a third group of high school students that was also afforded the opportunity to answer questions during initial learning of the lecture, but only after the final portion of the lecture. We have previously shown that students experience considerable difficulty answering questions after an extended period of study that does not involve interpolated testing (Szpunar, McDermott, & Roediger, 2008; Szpunar et al., 2013), and sought to assess whether this salient experience with difficult-to-answer questions would help students to lower unrealistic judgments of learning.

2. Method

2.1. Subjects

Fifty-four high-school students (ages 16–18 years) attending Harvard University's summer school program participated in the study. High-school students were recruited because they had little-to-no prior experience with statistics. Students provided informed written consent, obtained parental consent if they were younger than 18 years, and were randomly assigned to one of three experimental groups.

2.2. Study materials

An introductory statistics lecture was used in the experiment (Statistics 104, Department of Economics, Harvard University). The 21-min video covered basic introductory concepts in statistics (e.g., outlining the relation between a sample and population) that did not require past experience with statistics. The video was divided into four 5-min segments using iMovie software (Apple).

2.3. Design and procedure

Students took part in a 1-h learning session. Students were told that the video-recorded lecture would be divided into four segments of equal length. Further, students were told that they would complete a number of tasks between each segment. Initially, students were informed that they would complete 1 min of math problems after each segment of the lecture that was unrelated to the content of the lecture (six problems were presented and students were given 10 s to answer each problem; e.g., $12 \times 7 = ?$). Moreover, students were told that either two more minutes of math problems (12 problems; 10 s per problem) or a 2-min quiz about the most recent segment of the lecture (six questions; 20 s per question; e.g., What is the relation between a sample and population?) would follow the first minute of math problems (following each segment). Note that the test questions were brief short answer questions that tapped memory for information explicitly presented in the lecture. Importantly, students were informed that a computer program would randomly determine the occurrence of the quizzes, such that students might experience 0–4 quizzes during the lecture. For example, students could be quizzed after each lecture segment, after none of the segments, or anywhere in between. Finally, students were told that regardless of the frequency of testing during the lecture that there would be a final cumulative test that would test their knowledge about the entire lecture. In reality, one-third of the students received tests after all four lecture segments (4-test group), one-third of the students received a test following the fourth and final segment of the lecture (1-test group), and one-third of the students did not receive any tests during the lecture (0-test group). After the lecture was complete, students were given a 5-min break during which they played an online computer game (Tetris). After the break, students were asked to predict, on a scale of 0–100%, how well they thought that they would perform on the final cumulative test. The final cumulative test included the same 24 questions that were presented to students in the interpolated group, and students were allowed to complete the final test in a self-paced manner. Note that students were not given any indication about the types of questions that they would receive on the final test. The lecture, math questions, and quiz questions were presented on a computer screen using E-Prime 2.0 software on a Dell desktop computer, and responses were made using a keyboard.

Finally, we set out to assess whether our previous demonstration that interpolated tests helped students to avoid mind wandering and engage in note taking (Szpunar et al., 2013) would extend to a population of high-school students. In order to do so, our experimental design also incorporated the following features. First, students were told that a visual cue indicated by the phrase "Mind wandering? Yes/No" would appear on the computer screen at some random points during the lecture, and that whenever they saw this cue that they should respond on a sheet of paper by writing the word 'yes' or 'no'. The visual cue remained on the screen for 5 s as the lecture continued, and was accompanied by an auditory cue (i.e., a bell) that sounded during the first of the 5 s to ensure that students noticed the cue. Four mind wandering probe sequences were used in the study. For each sequence, the mind wandering probe occurred at a randomly determined time point during each segment that was at least 30 s into the segment and 30 s before the segment was complete. For instance, one sequence involved probes that occurred 96 s, 285 s, 201 s, and 155 s into the first, second, third, and fourth segments, respectively. The presentation of these mind wandering probe sequences was counterbalanced across students. Second, students were provided with the lecture slides associated with the lecture, and instructed that they could use the slides in any way that they thought might help them learn from the lecture. Upon the completion of the lecture, we retrieved each student's lecture slides. As a rough measure of student engagement, we checked to see for what proportion of slides students took additional notes, and whether interpolated testing influenced note-taking behavior. Note taking was defined in a manner such that both additional notes associated with lecture content and emphasis given to existing notes (e.g., circling or underlining key lecture points) were counted as additional notes. However, markings unrelated to lecture content (e.g., doodles) were excluded from the analysis.

3. Results

3.1. Initial test

In order to determine that students in the 1-test group were in fact making predictions of final test performance following an

Table 1
Mean (standard deviation) predicted and actual performance for high school students in the 4-test, 1-test, and 0-test groups.

	Predicted performance	Actual performance
4-Test group	77%(9.8%)	75%(15.5%)
1-Test group	60%(20.8%)	50%(15.2%)
0-Test group	78%(12.0%)	48%(17.7%)

especially difficult test during initial learning, we compared initial recall of the fourth segment of the lecture between students in the 1-test and 4-test groups. Indeed, students in the 1-test group (47%) performed considerably worse than students in the 4-test group [73%; $t(34) = 3.82$, $p = 0.001$, $d = 1.27$].

3.2. Final test

3.2.1. Predicted and actual performance

One-way ANOVAs revealed significant effects of testing on predicted [$F(2, 51) = 7.87$, $p = 0.001$, $\eta_p^2 = 0.236$] and actual [$F(2, 51) = 15.54$, $p < 0.001$, $\eta_p^2 = 0.378$] performance. With regard to predicted performance (Table 1), planned comparisons showed that whereas students in the 4-test (77%) and 0-test (78%) groups did not differ reliably in their predictions of final test performance ($t < 1$), students in the 1-test group (60%) made significantly lower predictions than students in the 4-test [$t(34) = 3.16$, $p = 0.003$, $d = 1.12$] and 0-test [$t(34) = 3.04$, $p = 0.005$, $d = 1.05$] groups. With regard to actual performance (Table 1), planned comparisons revealed that students who had been intermittently tested during the lecture (4-test group; 75%) outperformed students in the 1-test [50%; $t(34) = 4.84$, $p < 0.001$, $d = 1.62$] and 0-test [48%; $t(34) = 4.87$, $p < 0.001$; $d = 1.63$] groups. Students in the 1-test and 0-test groups did not differ from one another in terms of actual performance ($t < 1$). Critically, testing served to reduce overconfidence [$F(2, 51) = 8.40$, $p = 0.001$, $\eta_p^2 = 0.25$] such that students in the 4-test ($M = 2.67$ points overconfident) and 1-test ($M = 10.28$) groups were less overconfident than students in the 0-test group ($M = 29.83$), $t(34) = 4.56$, $p < 0.001$, $d = 1.55$ and $t(34) = 2.54$, $p = 0.016$, $d = 0.85$, respectively. Students in the 4-test and 1-test groups did not reliably differ from one another in this regard, $t(34) = 1.12$, *ns*.

3.2.2. Calibration

In order to further assess the extent to which testing improved calibration between predicted and actual performance, we calculated calibration scores – agreement between predicted and actual final test performance – using the method described by Miller and Geraci (2011), which transforms absolute differences between predicted and actual performance into a score ranging from 0 to 100:

$$\left(1 - \frac{|\text{Predicted} - \text{Actual}|}{100}\right) \times 100$$

Using this formula, a score of 0 represents complete inaccuracy and a score of 100 represents perfect calibration. As an example, a student predicting a score of 90% but achieving a score of 40% on the final cumulative test would have a calibration score of 50, whereas a student predicting a score of 90% and also scoring 90% would have a calibration score of 100. A one-way ANOVA revealed a significant effect of testing [$F(2, 51) = 7.10$, $p = 0.002$, $\eta_p^2 = 0.22$]. Planned comparisons showed that students in the 4-test group (89) were better calibrated than students in both the 1-test [78; $t(34) = 2.60$, $p = 0.014$, $d = 0.89$] and 0-test [69; $t(34) = 3.72$, $p = 0.001$, $d = 1.28$] groups. Interestingly, although the single test at the end of the lecture served to reduce overconfidence (see above), students in the 1-test group were not significantly better calibrated than students in the 0-test group [$t(34) = 1.43$, *ns*].

3.2.3. Mind wandering and note taking

One-way ANOVAs revealed a marginal effect of testing on mind wandering [$F(2, 51) = 2.94$, $p = 0.062$, $\eta_p^2 = 0.103$] and a significant effect of testing on note taking [$F(2, 51) = 5.51$, $p = 0.007$, $\eta_p^2 = 0.178$]. With regard to mind wandering, students in the 4-test group (22%) mind wandered in response to fewer probes than students in both the 1-test (39%) and 0-test (40%) groups, however, the difference was only reliable for the latter comparison [$t(34) = 1.98$, $p = 0.068$, $d = 0.64$ and $t(34) = 2.48$, $p = 0.018$, $d = 0.83$, respectively]. With regard to note taking, students in the 4-test group (68%) took notes for a greater proportion of lecture slides than students in both the 1-test [38%; $t(34) = 3.91$, $p < 0.001$, $d = 1.32$] and 0-test [48%; $t(34) = 2.15$, $p = 0.038$, $d = 0.75$] groups. The 1-test and 0-test groups did not differ from one another in either regard (t's < 1).

3.2.4. Additional analyses with college students

In order to further assess the reliability of the finding that interpolated tests served to improve calibration by acting upon actual and not predicted performance (i.e., 4-test group versus 0-test group), we re-analyzed/collected additional data pertaining to college undergraduates. To determine an adequate sample size per group, we conducted a formal power analysis using the effect size ($d = 1.28$) for the difference in calibration scores between the 4-test and 0-test groups of high-school students. This analysis indicated that using a sample size of 12 students per group would allow us to detect effects of interpolated testing on calibration accuracy with power equal to or greater than 0.80. For the 4-test group, a subset of 12 students were randomly selected from a prior dataset (Szpunar et al., 2013) in which college students had learned the same statistics lecture under the same circumstances as the high-school students in the 4-test group in the present study (i.e., the college students had also made a global judgment of learning). For the 0-test group, data from 12 college students were collected anew under the exact same conditions as high-school students in the 0-test group in the present study. There were no differences in the predictions of final test performance made by college students in the 4-test (83%) and 0-test (84%) groups ($t < 1$). However, students in the 4-test group (92%) outperformed students in the 0-test group [62%; $t(22) = 5.43$, $p < 0.001$, $d = 2.58$]. Importantly, students in the 4-test group (89) were significantly better calibrated in their predictions of final test performance than students in the 0-test group [76; $t(22) = 2.51$, $p = 0.020$, $d = 1.13$]. Although this additional analysis should be interpreted with some caution because the samples were collected at different time periods, the pattern of results was nonetheless highly similar to those obtained with the high-school students.

4. General discussion

The present study examined the extent to which testing could improve calibration between predicted and actual learning of a video-recorded statistics lecture. The results of this study are notable in three respects. First, students were generally overconfident in their learning of the video-recorded statistics lecture. Second, interpolated testing helped to bridge the gap between predicted and actual performance by improving learning of the lecture without producing a corresponding increase in predicted performance. Third, providing a single test for the final portion of the lecture, a condition in which learning is known to be especially poor (Szpunar et al., 2008, 2013), served to lower unrealistic judgments of learning. Taken together, the present results suggest that measures may be taken to improve calibration of learning from video-recorded lectures that target either predicted or actual performance. On the basis of our preliminary findings, it appears that

interpolated testing does the best job of fostering both a high level of predicted and actual learning.

We further demonstrated that interpolated tests helped high-school students to marginally reduce mind wandering and reliably increase note taking and retention. Although this pattern of data generally replicates our previous findings with college students, it is noteworthy that the interpolated testing intervention was not quite as effective in reducing mind wandering in high-school students as it was with college students (cf. Szpunar et al., 2013). Differences in interest level in the subject matter or method of delivering the mind wandering probes (e.g., college students in our previous study received mind wandering probes directly from an experimenter present in the testing room) are a likely explanation for this pattern of results. Alternatively, although speculative, it is possible that group differences in executive control may underlie the ability to reap the benefits of interpolated testing on attention to lecture content (Luciana, Conklin, Hooper, & Yarger, 2005). Future studies will be needed to distinguish between these various possibilities. With regard to the influence of interpolated testing on note taking and retention, it is important to note that interpolated testing may benefit retention of lecture materials via the influence of retrieval practice per se, an associated boost in note taking, or both. Future studies designed to more clearly tease apart the influence of testing and note taking on retention of lecture content should be highly informative.

5. Practical implications

Online learning is growing rapidly and playing an increasingly prominent role in both high school (Picciano et al., 2012) and college (Allen & Seaman, 2007) education. Video-recorded lectures represent one key component of learning in online environments (Breslow et al., 2013). Our study set out to assess the extent to which testing can be used to help students overcome the tendency to be overconfident in their judgments of learning associated with video-recorded modules. Notably, interpolating a video-recorded lecture with brief memory tests helped to boost learning in a manner that better calibrated actual with predicted performance. In the case where video-recorded lectures are not interpolated with tests, we showed that a test at the end of the lecture served to lower unrealistic judgments of learning. Moving forward, studies addressing the timing of interpolated tests, the role of exposure to questions/explicit feedback and re-exposure to lecture content, and the effectiveness of interpolated activities other than tests will be needed to better understand how structuring of video-recorded lectures can improve learning and meta-comprehension in online education.

Conflict of interest statement

The authors declare that they have no conflict of interest.

Acknowledgment

Supported by a grant from the Harvard Initiative for Learning and Teaching (K.K.S. and D.L.S.).

References

Allen, I. E., & Seaman, J. (2007). *Online nation: Five years of growth in online learning*. Needham, MA: Sloan Consortium.

Bol, L., & Hacker, D. J. (2012). Calibration research: Where do we go from here? *Frontiers in Psychology*, *3*, 229 (Article 29).

Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., & Seaton, D. T. (2013). Studying learning the worldwide classroom: Research into edX's first MOOC. *Research & Practice in Assessment*, *8*, 13–25.

Choi, H. J., & Johnson, S. D. (2005). The effect of context-based video instruction on learning and motivation in online courses. *American Journal of Distance Education*, *19*, 215–227.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, *22*, 271–280.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes*, *105*, 98–121.

Gal, I., & Ginsburg, L. (1994). The role of beliefs and attitudes in learning statistics: Towards an assessment framework. *Journal of Statistics Education*, *2*(2).

Hacker, D. J., Bol, L., & Keener, M. C. (2008). Metacognition in education: A focus on calibration. In J. Dunlosky, & R. Bjork (Eds.), *Handbook of memory and metacognition* (pp. 411–455). NJ: Lawrence Erlbaum Associates.

Luciana, M., Conklin, H. M., Hooper, C. J., & Yarger, R. S. (2005). The development of nonverbal working memory and executive control processes in adolescents. *Child Development*, *76*, 697–712.

Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2010). *Evaluation of evidence-based practices in online learning: Meta-analysis and review of online learning studies*. U.S. Department of Education.

Miller, T. M., & Geraci, L. (2011). Training metacognition in the classroom: The influence of incentives and feedback on exam predictions. *Metacognition and Learning*, *6*, 303–314.

Picciano, A. G., Seaman, J., Shea, P., & Swan, K. (2012). Examining the extent and nature of online learning in American K-12 Education: The research initiatives of the Alfred P. Sloan Foundation. *The Internet and Higher Education*, *15*, 127–135.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval in long-term retention. *Trends in Cognitive Sciences*, *15*, 20–27.

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210.

Salomon, G. (1984). Television is easy and print is tough: The differential investment of mental effort in learning as a function of perceptions and attributions. *Journal of Educational Psychology*, *76*, 647–658.

Schraw, G., Potenza, M. T., & Nebelsick-Gullet, L. (1993). Constraints on the calibration of performance. *Contemporary Educational Psychology*, *18*, 455–463.

Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational Psychologist*, *26*, 207–231.

Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 6313–6317.

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1392–1399.