Lecture Evaluations by Medical Students: Concepts That Correlate With Scores

Aaron Jen, BA, Emily M. Webb, MD, Bren Ahearn, MA, David M. Naeger, MD

Abstract

Purpose: The didactic lecture remains one of the most popular teaching formats in medical education; yet, factors that most influence lecturing success in radiology education are unknown. The purpose of this study is to identify patterns of narrative student feedback that are associated with relatively higher and lower evaluation scores.

Methods: All student evaluations from our core radiology elective during 1 year were compiled. All evaluation comments were tagged, to identify discrete descriptive concepts. Correlation coefficients were calculated, for each tag with mean evaluation scores. Tags that were the most strongly associated with the highest- versus lowest-rated (> or < 1 SD) lectures were identified.

Results: A total of 3,262 comments, on 273 lectures, rated by 77 senior medical students, were analyzed. The mean lecture score was 8.96 ± 0.62. Three tags were significantly positively correlated with lecture score: "interactive"; "fun/engaging"; and "practical/important content" ($r = 0.39$, $r = 0.34$, and $r = 0.32$, respectively; all $P < .001$). More tags (n = 12) were significantly negatively correlated with score; the three tags with the strongest such correlation were: "not interactive"; "poorly structured or unevenly paced"; and "content too detailed or abundant" ($r = -0.44$, $r = -0.39$, and $r = -0.36$, respectively; all $P < .001$). Analysis of only the highest- and lowest-rated lectures yielded similar results.

Conclusions: Several factors were identified that were strongly associated with lecture score. Among the actionable characteristics, interactive lectures with appropriately targeted content (ie, practical/useful) were the most highly rated.

Key Words: Medical students, medical education, lectures, teaching, feedback, evaluations

J Am Coll Radiol 2016;13:72-76. Copyright © 2016 American College of Radiology

INTRODUCTION

The didactic lecture is one of the most time-tested and popular teaching methods used today. As a simple method of transferring knowledge from instructor to student, its benefits are many. The format is scalable, allowing for instruction of tens of students up to thousands, particularly when information is disseminated via the Internet [1]. In addition, the format is very familiar, and can be time efficient for preparing and delivering content.

The traditional lecture format, however, is prone to several pitfalls [2]. Students' attention spans for passive learning are extremely limited, often to just 20 minutes [3]. Lecturer styles of presentation are unique, and sometimes ineffective, requiring learners to adapt to each presenter. Additionally, lecturers choose the level of information and the pace of presentation, which may not be ideally suited to all learners. Finally, lectures are limited in what they can teach; they often focus on facts and concepts, rather than on skills or creativity [4].

Due to these challenges, some educators have argued that alternative teaching formats should have a greater role in education. Problem-based and team-based learning are implemented widely in current medical school curricula [5-7]. Blended and "flipped" learning models are popular [8-10]. Particularly now, in the digital age, online interactive modules increasingly are being promoted.

Each method offers its own unique advantages and disadvantages, yet none has completely replaced traditional lecturing. For that reason, educators should continue to improve the effectiveness of lecture delivery. Efforts to improve lecturing have come far in recent years. Several authors have recently shared insightful techniques to promote active learning, including incomplete outlines,

Department of Radiology and Biomedical Imaging, University of California, San Francisco, San Francisco, California.

Corresponding author and reprints: David M. Naeger, MD, Department of Radiology and Biomedical Imaging, 505 Parnassus Ave, M-391, Box 0628, San Francisco, CA 94143-0628; e-mail: david.naeger@ucsf.edu.

The authors have no conflicts of interest related to the material discussed in this article.

1546-1440/15/$36.00 ■ http://dx.doi.org/10.1016/j.jacr.2015.06.025

break activities, relatable examples, and teachable skills [4,11-14].

Kessler et al [15] compiled presentation techniques from highly rated faculty at the American College of Emergency Physicians, focusing on key objectives, increased audience participation, and manageable slide content. Multimedia design principles have been applied to lecture creation. Issa et al [16], for example, found that audiences preferred a lecture design focused on visual representation, and without bullet points. To our knowledge, very little of the radiology literature has examined what characteristics are most desired by learners, particularly medical students.

In this study, we analyzed medical student feedback on lectures in our core senior radiology elective. Specifically, we analyzed all narrative comments collected during 1 year, and determined the comment phrasing that was most associated with various average numeric evaluation scores. Our hypothesis was that high- versus low-scoring lectures would be associated with a definitive set of unique adjectives.

METHODS

This study is exempt from institutional review board requirements. Only anonymized lecture comments and scores, extracted from an existing course-evaluation database, were reviewed.

Lecture Evaluations: Scores and Comments

Feedback is routinely obtained for all lectures in the core senior student radiology elective at our institution. Anonymized scores and comments from all iterations of the course taught in 2014 were included in this retrospective analysis. Each lecture's numeric score (range: 1 [worst] to 10 [best]) was an average rating from all students in attendance; the free-text comments consisted of all written comments. Providing both a numeric score and a written comment (even if just one word) was required. The actual content of the evaluations, which were originally collected for the purpose of course improvement, was anonymous, a fact known to the students providing the evaluations and comments.

Data Extraction

Deidentified free-text comments for each lecturer were reviewed for specific adjectives or singular concepts, and semantically tagged by one investigator. A list of tags was generated based on the content encountered. If a sentence contained an adjective or concept that was the same as or synonymous with one tagged in a previously reviewed comment, it was tagged with the same label. If a concept was entirely new, it was added to the list. The final list of discrete adjective families contained 42 tags. In cases in which comments were difficult to tag, all the investigators reviewed the comments and decided on an appropriate categorization, by consensus. This process was needed for approximately 5% of comments.

Each anonymized student comment (a sentence, sentence fragment, or rarely, multiple sentences) for each anonymized lecture was given a "yes or no" value for each of the 42 tags. Most comments were only one sentence, so most tags were not represented by any given comment. Each lecture was assigned a percentage frequency for each of the 42 tags. For example, if a lecture was evaluated by 12 students, and 4 stated that it was "interactive," and 3 stated that it was "too long," then 33% of the comments were positive for the tag "interactive," and 25% were positive for the tag "too long."

Statistical Analysis

The correlation between lecture score (range: 5.7-10.0) and tag percentage (range: 0%-80%) was analyzed using the nonparametric Spearman rank-order correlation coefficient. To determine a statistical significance cutoff, a Bonferroni correction was applied to account for the many analyses. A total of 42 correlation analyses were assessed, so an overall significance level of $P < .05$ was selected and divided by 42; individual analyses with $P < .0012$ were considered statistically significant.

We sought to confirm the correlation analysis by analyzing only the most-extreme groups of lectures, specifically, the highest- and lowest-scoring lectures (as defined by a score >1 SD away from the mean). The frequencies of each tag for lectures in the highest- versus lowest-scoring group were compared using the Mann-Whitney U test. A Bonferroni correction was applied as well, so individual analyses with $P < .0012$ were considered statistically significant.

RESULTS

The feedback from 317 lectures given by 54 different lecturers was available in the course-evaluation database. Forty-four sessions were excluded from analysis, owing to nontraditional lecture formats (eg, orientations, tours, and hands-on modules). These formats differed fundamentally from the traditional lecture model, and could, theoretically, have been evaluated differently by the students. The final sample included 273 lectures given by 48 lecturers. These lectures were evaluated by a total of 77