How do course evaluations improve teaching and learning?

Andrew Ho, *Harvard Graduate School of Education*

Remarks prepared for a faculty luncheon: Talking about our Teaching
February 28, 2019.

Announcement: https://linc.seas.harvard.edu/event/faculty-luncheon-talking-about-our-teaching-0
Slides: https://drive.google.com/file/d/1m064r_fO88jxzZqe6fW0R83T5cobgYaf/view
Andrew Ho's HGSE Course Evaluations: https://scholar.harvard.edu/files/andrewho/files/s-052_course_evaluations_spring_2018_andrew_ho.pdf

Abstract:

We intend course evaluations to answer three questions, from three audiences, for three purposes. 1) From prospective students, for course selection, "What did my fellow students think about this course last year?" 2) From instructors, to improve their instruction, "What did my students think that I could do better next year?" 3) From administrators, to inform promotions and tenure, "Is this candidate fulfilling or excelling in the teaching mission of this school?" Through poor design, poor timing, and poor analysis, all three audiences are underusing and misusing data from course evaluations. I will review best principles from the design, reporting, and deployment of course evaluations. I also propose some pilot studies to improve the use of course evaluations for each audience, for each purpose.

Notes:

Thank you all for having me. As you've seen in my bio, I am a psychometrician, and I teach statistics and measurement at our Graduate School of Education. In my research, I work to improve the design and use of educational tests. One of the central concepts in my field is that of validity. Our use of the term is at odds with its colloquial use, where folks often say, "this test is valid and reliable." We never say that. In modern testing theory, validity is a property not of a measurement instrument but of intended uses of its data.

What does this mean for course evaluations? It means that no matter how well we design and deploy the items, nor how poorly, we would never say that "course evaluations are valid," nor that "course evaluations are invalid." We are trained to evaluate the use of each instrument for a particular purpose.

We must start validation, therefore, by stating the intended purposes. And what I find interesting about course evaluations, and perhaps you've noticed this as well: is that there are at least three distinct purposes that we intend course evaluations to serve. If there is one lesson we have learned through the history of my field of educational testing, it is that efforts to use one instrument for multiple purposes never end well. The metaphor I imagine is a game that might have been created for that old show, *American Gladiators.* I imagine three players tied together by a giant rubber band, each trying to reach their own one of three goals arrayed in a triangle, beyond each of their reach. As any person tries to reach her goal, the rubber band limits the others from reaching theirs. The rubber band represents the limitations of a single instrument. The fact that advancing toward one goal means sacrificing progress toward another reflects the tradeoffs inherent in using a single instrument to serve multiple purposes. And in the case of course evaluations, the metaphor suggests that if one player were unhindered by others, she still couldn't reach her goal on her own.

My learning objectives for our discussion today is that you will leave being able to 1) identify these three purposes, 2) explain why using course evaluations, alone, to serve any one of these purposes (let alone all of them) is folly, and 3) leave committed to trying out at least one pilot study to supplement course evaluations in your own teaching or your own department. My current belief is that course evaluations should be improved and supplemented, not discarded nor dismissed. I look forward to convincing you that this is true and working together to improve and supplement course evaluations. Or I look forward to you convincing me that perhaps I am wrong or naïve.

As you read in my abstract, I believe that course evaluations answer three questions, from three audiences, for three purposes. First, prospective students ask, "What did my fellow students think about this course last year?" Second, instructors ask, "What did my students think that I could do better next year?" Third, administrators ask, "Is this instructor fulfilling or excelling in the teaching mission of the school?" I will address each of these purposes in turn.

The Student Audience: What did my fellow students think about this course last year?

I'll start by discussing how course evaluations answer prospective students' questions about what the previous year's students thought about the course. The theory of action is that students will make more informed choices about their courses, and that this in turn improves their learning, their enjoyment, and other future outcomes. The most relevant question on the course evaluation is, for example, on the GSE course evaluation form, Question 9: "What advice would you give to students who are thinking of taking this course (about its level, the amount of work required, any prior training needed, ways to get the most out of the course, etc.)?"

As you can see, another theory of action that Question #9 reveals is not about course selection but instead that course evaluations contain good learning tips. They can reveal the "hidden curriculum" of the course. I spend half of my first lecture in statistics trying to lay this hidden curriculum bare. I tell students that most of their learning happens not in lecture but outside of class, in sections and with each other. I tell them that they should print assignments the day they are released and actively consider the questions as they are learning about the relevant content in lecture and section. I tell them that this is a writing class more than it is a math class, and they should choose their words, when they interpret results, with great care. These are tips that I learned when I was a student in statistics courses. My professors never told me these tips outright, and I wish they had, so I do so now in my own classes. Similarly, feedback from fellow students can improve not only students' selection of courses but also improve their learning strategies for courses that they select.

As I noted, I believe that course evaluations should be improved and supplemented, not discarded nor dismissed. How can we improve course evaluations for the purpose of informing students about course selection? First, course evaluations should not be a sole deciding factor. Other tools, some of which our own Office of the Vice Provost for Advances in Learning (VPAL) are piloting, may help to inform students about the substance of the course, its relevance to their interests, and its importance on a pathway to future courses and careers, as a supplement to the opinions of past students. No single instrument can should serve this purpose alone. Second, and this is a point to which I'll return, there is compelling evidence that course evaluations reflect the biases that students have against women, against underrepresented minorities, and against younger instructors. If students select courses on this basis, they do so against the interests of their learning.

Third, transparent communication about workload and grading increases a risk what our colleague Chris Winship has called “the low-low contract”: instructors hold students to low standards by giving them high grades for little work, and students reward the instructor with high evaluation scores. This signals to future students that this is an “Easy A” and perpetuates a cycle that subjugates learning in favor of fake praise. In one interesting pilot study at Stanford, they found that revealing the empirical grade distribution of past students (which was, like Harvard, pretty skewed with many As and A-s) hurt the outcomes of students, perhaps because they believed that such a distribution was automatic and not the result of hard work. In this way, evaluations can allow students to seek out paths of least resistance.

In spite of these flaws, I believe information that students provide in items like Question #9 are, on balance, useful to their prospective student audience. I think they help to signal that we have a community where we can talk about good teaching and learning openly. One strong recommendation I have for a pilot study is for you to include a question like Question #9 in your own evaluation instruments, if you do not already. Another is to foreground this item in reports or dashboards of evaluations that are designed with the prospective student audience in mind, before other items which serve other purposes. A third suggestion is to frame the information to make students wiser readers of evaluations. They should know, for example, that any large course is likely to have outlying comments that should probably be taken with a grain of salt. And a fourth suggestion is to allow the qualitative responses to this item to be ranked or flagged for usefulness. Imagine an Amazon-style rating system where folks could vote on the usefulness of comments. In addition, instructors could choose comments that they found most informative for prospective students, and these could be labeled as “instructor-selected.”

The Instructor Audience: “What did my students think that I could do better next year?”

One of my favorite prompts for getting constructive feedback is the “plus delta,” a paired prompt that leads with “what did you like?” (the “plus”) and continues with “how can we improve?” (the “delta,” the Greek letter that often refers to change). Most good course evaluations are structured in this way, where the items encourage respondents to focus on constructive suggestions. The GSE course evaluation asks, “In what ways was the instructor most effective? Why?” and continues with “What recommendations would you make to the instructor to strengthen his or her teaching and/or make the course more valuable?” The theory of action here is that students know what can improve their learning; they articulate this in their evaluations; instructors will read these suggestions; and they will respond accordingly the next year.

In testing we distinguish between summative and formative assessments, where summative assessments are valuable summaries but usually too late to be actionable and make a difference. Here, too, course evaluations are so obviously ineffective at improving instruction in the current semester that the act of requesting a course evaluation can seem like an insult to the student. They might respond, “only now, after it’s over, you ask me how I think you can improve?” There is an array of methods that we can use to supplement end-of-semester course evaluations with student feedback during the semester. Here are just a few, some well-established, others that I’ve found that work for me.

First, the exit ticket, a quick survey that folks have to submit to “exit.” Just go to forms.google.com and create a bit.ly link, or for small classes print and collect the scraps of paper themselves. Either do a “plus-delta” prompt for pedagogy or a more conceptual and substantive plus-delta, “what did you learn? / what could we spend more time covering next time?” Exit tickets are a well-established technique for any instructor’s toolkit.

I personally find exit tickets too awkward to process effectively on a regular basis. Instead, I use three separate approaches. In class, I use what I and my colleagues call "Quiet Questions" and "Learning Checks." Quiet Questions are an open-edit google document into which students can type any questions they like, live in lecture.[1] My TFs staff this during class and respond to questions in the moment. Students may always ask questions out loud in lecture, but the Quiet Questions document allows students who are shy or who may have what they think is a trivial question to get a quick answer. The TFs may encourage the student to ask the question out loud if it is a particularly good question or if many students share the same question. In addition to providing a parallel outlet for students to ask questions in class, Quiet Questions also serves as a kind of "exit ticket" for me. I review the Quiet Questions after each class and elaborate upon the TF responses as necessary. I can use this review time to evaluate how the lecture went and what concepts I think I should review in the next class. Meantime, the Quiet Questions become a kind of running set of class notes for all students in the class. You can see my current Quiet Questions at http://bit.ly/s52stats19archive

Learning Checks are standard in-class quizzes. I run them using the "think-pair-share" protocol, where I have students think about the answers on their own, then talk the answers through with a neighbor. This lets them get to know each other and helps them get used to talking about statistics out loud. I use google forms for this, and it helps me to assess how students are internalizing the material while I'm teaching it.

Finally, perhaps most importantly, I follow my colleague Judy Singer's practice of convening a series of brown-bag, bring-your-own lunches throughout the semester, with 8-12 students per lunch. I do these by program, so each lunch has students from our international policy program, or our developmental psychology program, so they know each other and are comfortable. In the first half hour of lunch, I let them ask questions about anything EXCEPT the course itself, my academic history, my personal history, my research, my family, my pop culture recommendations, whatever. I take a few questions at a time and answer them together if it's crowded, so I can get as many folks actively participating as I can. Then, in the second half of lunch, I solicit plus-deltas from everyone and simply take notes. It's important to start with the open questions and not the plus-deltas, because plus-deltas will always fill an hour on their own. It's also important to require students to provide both plusses and deltas lest the vibe get too negative. I find these brown bag lunches are an excellent way to get feedback throughout the semester and get to know the students more personally.

What about the actual use of summative, end-of-semester evaluations, even as delayed as they are? If you are like me, you find these can sometimes be hard to read. A large proportion are positive and constructive, but a few are personal, ad hominem, unfair, and difficult to shake. They can make me feel cynical, frustrated, and depressed. They are not helpful. My colleague Josh Goodman at the Kennedy School recently tweeted about his solution, and I know Todd Rogers and others do something similar. The idea is simple: Have a colleague or, if resources permit, task an assistant to read the evaluations and divide them into three categories, as Josh describes: "positive feedback, constructive criticism, and unconstructive criticism." The colleague then makes a document that contains the first two and not the

[1] There is compelling evidence allowing laptops, tablets, and phones hinders learning, e.g., https://www.brookings.edu/research/for-better-learning-in-college-lectures-lay-down-the-laptop-and-pick-up-a-pen/ I do not allow laptops without special permission in my smaller fall seminar (15-20 students). For my larger spring statistics class, however, I have so far felt that my early admonishments help my students to monitor themselves and each other, and the benefits of quiet questions outweigh the potential cost of distractions.

third. Then, you can do the same for your colleague, and get together for lunch to talk about it. If the goal is improving teaching and learning, I agree with Josh that there is no reason to dwell on or be distracted by unconstructive criticism.

The Administrator Audience: "Is this candidate fulfilling or excelling in the teaching mission of this school?"

Finally, it's time to tackle the elephant in the room, the use of course evaluations in promotion and tenure decisions. I alluded earlier to the growing evidence that evaluation scores are biased against women, underrepresented minorities, and younger instructors. Some of this evidence is convincingly causal, such as in a study of course assistants in an online course who randomly assigned themselves different online gender identities (this and similar evidence is reviewed by Inside Higher Education here and here). I believe that course evaluations reflect existing biases of students, and I am not so naïve that I don't think Harvard students are biased. There are other biases one might expect, too, such as those for required courses, or larger courses, or courses that assign lower grades. If course evaluations are biased, how could I possibly recommend their continued use for high-stakes decisions like remediation, promotion, and tenure?

The solution in my mind is no different than that for educational testing in my own field: multiple measures, evaluated by professionals. Just as a single test score should not solely determine admission to a college, course evaluation scores should not form the whole (nor any substantial part) of a high-stakes evaluation of teaching. At the Ed School, we provide these so-called dot plots of our quantitative ratings (see below), complete with norms for similar classes and interquartile ranges to illustrate variability. But these are only a piece of a portfolio that includes full qualitative comments from students as well as a teaching statement. Our Promotion Committees summarize these in a report, and the raw data are available in portfolios. Do they weigh heavily in tenure decisions? Usually, no. Still, I believe the consensus that the prominence of teaching in tenure (and hiring) decisions has grown. And I also believe that its prominence is still not great enough.

I propose for your consideration a couple of next steps to conclude this talk. First, I think that more institutional research should be invested in simple descriptive analyses of course evaluation data to describe relationships between important factors known to bias or predict differences in course evaluation scores, including gender, underrepresented minority status, instructor age, enrollment, grade distribution, year of study, and required status. We should set norms appropriately in our interpretation of quantitative evaluation scores rather than comparing apples and oranges.

Second, my more audacious proposal is to evaluate teaching for tenure the same way we do research: with named comparands. For promotion and tenure decisions, we could identify 2-3 internal instructor comparands for the candidate. Rather than pulling selected quotes from an internal candidate's evaluations or comparing average scores to abstract dots on dot plots, these comparands would be specific instructors and their courses, including members of the senior faculty, chosen for illustrative similarities or differences. I imagine my comparands would be my colleagues John Willett, Judy Singer, and Dan Koretz. We could use not only dotplots and evaluations but other data sources as necessary, including course videos. The committee's recommendation would include a critical comparison of the candidate's quantitative and qualitative evaluation data and other teaching artifacts to those of named comparands, with careful attention to the programmatic context, the student composition, and the biases that may be affecting the ratings. This holistic perspective on teaching would be a great improvement to simple scores with limited context and known biases.

Reflections

There are a few additional topics that I consider beyond the scope of this 15-minute talk, but I think they are important to raise here and tackle elsewhere.

First, what do we do about the fact that course evaluations may not correlate with “value added” measures for an instructor’s contribution to a student’s actual learning? There is a whole field on this topic in K-12 education to which I and my colleagues at HGSE have contributed. Unfortunately, the assessment challenge in higher education is immense. Few subjects have the size nor the structure to provide high-quality standardized assessments across instructors. There has been some good work on this done in Physics and I look forward to seeing more. However, I do think that course evaluations still have a place. We should care not only that our students have learned but that they believe they have learned. We should care not only that our students have learned but that they are enthusiastic about learning more. I think course evaluations have a place beside test-based learning measures. And I hope these are both supplemented by other rich data sources about the instructors who close racial and gender gaps, in learning and in course-taking patterns, in choices of majors, in graduation, in employment, in earnings, and in flourishing. We should link our rich data sources in admissions, academics, advising, and alumni offices to learn more.

Second, what about items and the general design of evaluations as measurement instruments? Again, I think it is much more important to get feedback during the semester itself, rather than the end. But we can improve the design of summative course evaluations, too. There is again much to say here, and I refer you to the HGSE instrument as one that I think is generally designed well but, like all instruments, could be improved. I think instruments should have both quantitative and qualitative items and should generally be cut in half or a third in length. Instruments are too often designed by committees who do not have the three core purposes clear in their charge. Items get added because someone thinks “it would be interesting to know X,” not because they solve a particular problem. This results in bloated instruments that waste everyone’s time.

Third, I hope we can do more research at the specific moments that these evaluations (and other data) shape perceptions. These kinds of cognitive interviews, with students as they read evaluations, with instructors as they review their own, and with tenure review committees as they review evaluation data, are in my experience never done. These are standard steps to improve validity in the development of any instrument. Why not take them?