What Are Students Learning, and From What Activity?

David E. Pritchard
Cecil and Ida Green Professor of Physics
MIT Department of Physics
Developer of Mastering Physics

I am interested in figuring out how students learn, what they should be learning, and how we can use technology to increase the amount students learn by as much as possible. Let me start off by first quoting education researcher, Jill Larkin, who said education reform is a lot like farming. At the end of the season, you look at how big the harvest is, and then try to figure out what on earth you did to make it bigger or smaller than last year. Of course it might well be the weather, which you didn't have anything to do with. Education reform is like this because if you only measure at the beginning of the year and the end of the year, how do you know what course element caused the learning that happened?

To get some indication of what course element causes learning, we did some correlation studies, looking at the remedial physics course that I was teaching for the kids who hadn't done well enough to get a C in the fall. We looked at what correlates with their final exam in May being better than the one they took in December (this difference is what we define as learning). What activities did they do? It turned out that going to class didn't help much, statistically. The number at the bottom is a p value - the chance that the result that we obtained can be explained purely by chance. If $p > 0.05$, most people regard the result as “not statistically significant”. The group problems helped one year, but not the other. But the Mastering Physics was the absolute killer. The more the kids did of this, the more their final exam grades improved. Mastering Physics is the commercial version of the program that my son and I developed that was initially called cybertutorlmit.edu. It's a Socratic tutor. I can't go into detail about it in the time that I have. But as you can see, its use correlates with a two standard deviation learning effect, which is huge. Most classes do about one standard deviation over the whole term.

We then started looking in even more detail at the log of all the student’s interactions with this tutor. Let me emphasize a point of great educational potential: when you are doing online education, your system can generate a pile of data about learning like you've never had before. When I am sitting down at the end of the term assigning grades in a conventional course, I have my grade book: 12 homework assignments, four of them partly copied; three exams; and a couple of lab reports, definitely both copied. Based on those 100 bytes of information, I am trying to decide something about the learning of that student. When somebody finishes with Mastering Physics for a term, I have a quarter megabyte of data. I am overwhelmed. But Educational Data Mining is giving some novel insights into what’s happening.

I have some other research that I don’t have time to discuss here that really shows the importance of good habits on learning. Now I'll show you one about bad habits on learning. We started to look at “how long does it take students to finish the problems?” in the online environment. Time to do a task is a variable frequently studied by experimental psychologists with both human and pigeon subjects. We discovered a small percentage of students who could answer some questions requiring two or three analytic responses in 30 or 40 seconds! So we had a conundrum; “Okay, this problem took most of the kids ten minutes, and they made mistakes. How did these few answer it in less than a minute and not make any mistakes?” Well, of course, there are a lot of geniuses here at MIT, but a more likely explanation, we decided, was that they were copying their answer from somewhere. So this is the fraction of electronic homework that was copied. We grouped the students into a few groups on the basis of the overall percentage of their problems that they copied: less than ten percent here, and the students who copied 70 percent of their online homework way out here.

How did they do on the final? We found two different things. For the analytical problems on the final, we found this group of points here. These data are almost low enough error for me to publish if they were from my atomic physics lab! They are really outstanding from the point of view of education. If you know about p values, and I did not write it on here, but the p value here is 10 to the minus 15. So what you are seeing is not a statistical fluke.

On the other hand, on the conceptual questions that they learned in class (or from easy Mastering problems early in the assignment and in the term which they generally didn’t copy), the learning was independent of how much they copied on the homework. This is a huge differential effect in learning. It strongly suggests that all our students can learn physics if they put their mind to it, and that when they don't do the homework, it really costs them. Homework copying is over three times better at predicting a poor final than is doing badly on the pretest, or the first midterm, or on the homework score. Doing the work trumps native ability!

That is the kind of detailed information and investigation that we can find from educational data mining.

Next we decided to look at “What does two sigma learning really mean?” We decided to look at what the students were learning by comparing two groups two sigma apart. This is the final exam histogram, comparing 60 students in this group one sigma above average, and 60 students in that group, one sigma below. We are basically asking, “What do A- students know or do that C students don’t?”

Here is the quality of the analytic answers that they gave. i is “horrible,” basically “no clue.” ii is “Some inkling when you start off, and then you made some serious mistakes.” iii is “This was a so-so attempt, but your mom would be happy.” iv, “You got it pretty good,” and v is “You did it right— your dad would be happy.” The C students are spread all over this map, but maximized at “clueless.” The A students are up here at “getting it done.” There is a huge difference, when you look at the quality. Six times more A students can do the problem really well, and only a fifth as many are clueless. The majority of the C students wouldn't even make their Mom happy even though we are passing those C students. The results are similar for the written plan. Again the A- students are qualitatively far superior to the C students. I won't go into that further.

Now let's look at the scores. We find that the C students receive exam scores that are 38% lower than the A's and only 19% below average, and only 9 points lower than average when the test is graded on the basis of 100 points total. These statistics strongly minimize the great qualitative gaps between the A – and C students. The bottom line of all this is that partial credit grading, which is what we do, rewards partial understanding (see slide #7). Take it or leave it. We've got to look at what our system really does before we can reform it. There is a pretty clear challenge here: are we content to pass kids who have less than one chance in six of demonstrating a fairly complete solution to the questions on the final?

Now I'd like to address a second issue that has always worried me. We decided to look at it carefully. That issue is, "What do seniors remember from freshman physics?" If it is going to do them any good in the world after graduation, they at least ought to know this material when they collect their diplomas, right? We got some seniors — we had to pay them— who were hanging around the week before graduation. They took basically an equivalent freshman physics final exam, and we looked at how they did. We also administered some other tests and surveys.

How do you judge the learning or forgetting when you have two tests? This is really a good way to study education. What you do is you look at the students on the basis of their first test— the pre-test, as it is called. Then you plot how much they changed from there to the second test. If they got a zero on the first test, they could get 100 percent on the second test, and they could learn 100 percent. If they got 80 percent on the first test, then they can only learn 20 percent. This line here represents getting 100 percent in the final, and therefore that you learned 100% of what you didn't know on the pretest. So any straight line that goes through the point zero gain at 100% pretest like this is what we call a pure learning curve or pure gain curve. It means that everybody learned a constant fraction of what they did not know on the pretest. That turns out to be a pretty good model in lots of cases. It isn't understood why. We published a paper suggesting that it is a result of pure memory learning, which didn't go down very well with those of us who have the modern constructivist view of learning, which is most of my education colleagues.

On the final retest, after over three years without review, we expect that the students are going to do worse, not better. To treat loss of knowledge, we can look at the analog to pure learning, which is pure loss. You're on this line if you forgot everything from the first test— if you were up here at 80. Now, you forgot 80 percent, so this means you are forgetting everything. So a curve that passes through zero at zero and has negative slope like this would be a pure loss curve.

With this in mind, we gave them a conceptual physics test, and we were blown away. They did a little tiny bit better than they did at the end of freshman year. Wow! Then we looked carefully, and we found that actually there were two parts of the test. There was a part that was on basically math, mostly calculus and graphs. What we found there was that the graduating kids are up here. This is a fit to it. This is a pure learning, with a gain of 70 percent. At MIT they learned 70 percent of what they did not know when they walked in the door. A lot of them in fact got 100 percent on it. You can see that line there.

Then, at the end of the freshman year, we had data, too, on the same instrument. They had gained 35 percent. So they learned half of all they learned in math in first semester physics, or maybe some in the math course they were taking at the same time. Then they learned some more in the rest of their careers (probably mostly in second semester math and physics).

Well, that's the good news. Now I'll tell you the bad news. The remainder of the test is on physics concepts with some numerical physics calculations. We didn’t get a pure loss or gain curve when we plotted it against either the pre- or the post- freshman test scores. To get a clear loss curve, we had to plot it against how much they learned in the freshman course. What they learned in the freshman course, they forgot half of. They forgot, in fact, 52 percent. So we are seeing that students forget 50 to 60 percent over the four years. This is true for the analytic problems on the final as well – we get about 60% loss.

What that means is that the forgetting process between freshman course and senior graduation turns the A's into D's. It is worse than the difference between A and C that I just showed you. The former A kids are now D's. They would not pass the course— or would not be allowed to go on— even though they had an A as freshmen. If they had a B before, they would get an F. That is how much the forgetting there is.

Our data are consistent with the adage, “Use it or forget it.” Whether that implies that we should only teach, or mainly teach, what will be reused later, I am not so decided on. But I definitely think we ought to work more on habits than on facts, and we ought to work on procedures also. I'll ask everyone who's a parent here, did you work a lot to get your kid to memorize the multiplication table or the capital of the state? Or did you try much harder to get them to have decent habits, like starting the homework at a reasonable hour, or picking up their homework papers and putting them in a pile, so when they got up in the morning they could just grab that and go to school? That's where all the hassle was in our family, and I think that learning and thinking habits is where we should put a lot of effort in college education.

The steep loss of unreviewed learning, and the persistence or growth of what was reused impact the old question of what we should be teaching in introductory physics. This epiphany came after I'd been working in education for seven years and teaching for 37. I finally said, “Well, maybe our current final exam is not the end all be all measure of what we should be measuring. Maybe we should ask some people what they think we should teach.” So I basically asked a lot of different experts, maybe 500 in total. “Given a change in the academic calendar, you have 20 percent more time to teach the calculus-based introductory physics course to non-physics majors. The syllabus has not been expanded, so what learning will you seek to add or emphasize with this extra time?" [Introductory physics is taken mostly by non-physics majors, especially at MIT.] ""

This graph shows you the 12 choices the experts suggested grouped into 4 categories. It also shows you where the teachers— and I had several different groups of teachers— came down. I want to draw your attention to a couple of points here. The first one is that all teachers are in accord that we teach too much content, and the last thing we should do is add more topics. Secondly, the number two choice of all these different groups, and the teachers' top pick on average, was sense making. Sense-making involves questions like: when your computer has told you to make 21-inch cables for the suspension bridge, how can you be really sure that that is the right answer? Or do you just believe the whatever the computer says? Of course not. So you have to learn to think about scaling, and ratios, and common sense. You compare that bridge with the Verrazano-Narrows Bridge, which it is similar to, and scale from that. That is the kind of thing that we think is really important to teach the students.

After I'd determined what the teachers think, I said, "Well, why don't we ask the students what they want to learn – we don't necessarily know best?" So we did, offering them the same set of alternatives as we offered the teachers. What we found were some pretty big discrepancies. First we found a lot of things where the votes averaged to around the "by chance" reading of 10 percent, where the students and the teachers picked the same thing. But there were some really large discrepancies. Here are the teachers, on "wider content". There are the students. What the students wanted most was sexy new topics.

Second point. Here are the teachers on sense making. That was their top choice. Here are the students on sense making. We did not teach them to make sense of anything in high school. We taught them to get the answer and go on to the next problem. Why would you look over your answer on a timed test, or on homework that constitutes only a small fraction of your grade but a large fraction of your sleep budget? That's their attitude. However it turns out that the students really want to see the relationship of what they are learning to everyday life. Obviously that is not happening in the labs, because they don't think the labs are very good -- even though the labs are supposed to be a confrontation of what we are learning with the real world. So things are pretty badly off kilter here. Basically, the conclusion I come to at the end is that we experts want to teach the students to become experts, whereas the students want to learn why physics is interesting enough to warrant their investing the time to become experts. In these required courses, I think the mismatch between what they want to learn and what we want to teach contributes an awful lot to the general malaise characterized by, "Well, we're just taking a required course to get our graduation ticket punched, doing the problem sets week by week. We'll copy if we have to."

In the remaining part of the talk, I want to step back and talk about schools and the digital age. A lot of this comes out of a great book, *Rethinking Education in the Age of Technology*, by Collins and Halverson. Let's look at education from the point of view of preparing students for the world they are going to live in. We teach just-in-case knowledge. We're going to teach you the capitals of all the states, just in case you need to know that. We're going to test you on it, too, so don't laugh! When are we going to do it? Well, ideally everyone goes to college to age 22. Schools go September to June, 8:00 a.m. to 3:00 p.m. – the summer is for forgetting. We group the kids by age. The instruction, especially in K-12, is done on paper, and the teacher's role is to be the subject expert and to be the source and gatekeeper of the information.

How does this compare with life in the digital world? Most of the factual knowledge we need is found "just in time". I don't look up the melting point of tungsten anymore in my textbook or handbook. I just type "melting point of tungsten," and bing, there it is on Google -- one of the top three entries. Usually I can see the number right there in one of the excerpts without opening any of them. Then "when do people learn?" Well, to keep up with this accelerating world, you have to learn on the job and for the next job – i.e. all the time. Learning happens on cell phones in the subway. Obviously, this world groups people by interest, by their attainment level and by their profession. The age matters only in second order. Finally, most new written work is done on computers.

If you look carefully at what schooling should accomplish in the digital age, you'll have to conclude that the contemporary school is truly inappropriate. We have to rethink the necessity of "just in case knowledge" in light of the ubiquity of online resources and networked experts in our students' future. The appropriate teacher should be a coach and a guide, instead of a sage on a stage -- as much as I enjoy that role, like most professors!

Let me go on. Here my thinking is pretty speculative, although I back it up with some data, but not enough. I think the first thing about college is that we want to guide students individually both online and in class. Full time personal tutors in real life are too expensive, but personal tutors online are great. That is really what Mastering Physics is: an online personal tutor for your homework problems. We should incorporate games, simulations, social software, and class— the whole nine yards. But we have to integrate it.

Why is integrating everything so important? The answer is: so you can assess everything each student does to find out where his or her mind is. Because when you want to teach a student, you want to be able to teach them what they don't know, in a way that they will be receptive to, at a level that they can comprehend without boredom. You cannot do that unless you know where their head is. That is what a personal tutor really does. The tutor asks a question and looks into the eyes of the person, watches them struggle with it, and decides what to ask next or what help to give. We have to be able to do that online, and we cannot do that now, because we don't have good enough assessments.

Largely, education now runs in what engineers call "open loop mode." At the end of the term we survey the students: "Did you like this class?" I mean, does that question tell you how much they learn? We have a lengthy college-wide course evaluation. There is one important question, "Did you learn anything useful in this course?" But our questions mainly ask about the old-fashioned formal course components: “Was the textbook clear?” “Were the lectures clear?” “Did the teacher use the blackboard well?” “Did the teacher use the web resources well?” “Was the course well organized?” “Were the tests fair?” Statistical analysis tells that all you really have to ask is “Did you like the teacher?” Everything else correlates strongly with that.

Did you learn anything? In fact, most universities and colleges don’t care much about whether their students learned anything. If they did they’d try to improve it. I am omitting the fact that a lot of individual faculty members care tremendously, which is one of the reasons I love MIT – but I’m talking about the institution. When was the last time that somebody from MIT watched my class, then came into my office, and said, “You know, Dave, I think you could do better teaching if you did the following”? It happened only once - when I was teaching 8.01 with Tom Greytak - and I've been here a long time. And it was at Tom’s initiative – not MIT’s. Our educational institution does not reward us for increasing the learning of our students. If I am a “good teacher,” I am a good teacher because the students like me, not because they learned a lot. There is no measure of whether or how much they learned (except for the concept tests we give before and after).

Let me show you one small example of universal assessment. Say I try to figure out where my class is on a chapter-by-chapter basis by looking at their average grade on the homework. I put in the statistical error bars here. But the error is bigger than that, because on this chapter, they did great; they got 85 percent. But I don't know if that was because the related questions were too easy, or because the students were really skilled. The kind of data that I have now provides little guidance.

So what do I have to do to get useful detailed assessment? As an example, I can analyze the detailed data from the online homework using item response theory. That is what ETS uses to grade tests. It separates the skill of the students from the difficulty of the items, allowing a determination of student skill independent of which test items they actually do. We did this by comparing the Mastering Physics data from one of our classes with data from the entire system. Here is the result. Two points: you'll notice the errors are much smaller. And, there is a very consistent overall trend, not visible in the percentage correct graph. MIT students start 1.5 standard deviations above the national average, and then in about six weeks they decay to about half a standard deviation above the national average. So we see that our admissions committee selects people who are outstanding in knowledge when they come in the door, but not necessarily good at learning. Of course, the education system does not measure whether you are good at learning, so how can they tell?

Most of these points have error around a tenth of a standard deviation. That is about one point on a 100-point test. Now I may see that my students are a little weak on some topic, so I vary how I teach the material in that chapter next year. I can watch whether that point goes up or down. I can actually evaluate my teaching at least on the week-by-week basis, rather than on a whole semester basis. This example has given me as much information as if I gave a two-hour test every week. But of course, then I wouldn't have time to teach them very much. If you do integrated online education, you can get this kind of detailed high signal-to-noise ratio measurement for free, just by analyzing the data that is already in the system.

The other thing that is very important about having an integrated online system is that it allows improvement of the content through data mining. Before I give you an example, let me put this in perspective. You are teaching a class. You teach something. You give the kids a problem, and 80 percent of them get it, but 20 percent do not. How important would something be if it could reduce the fraction of the kids who were unable to solve it to 10 percent instead of 20 percent? That would be a pretty fantastic gain, wouldn't it? If we could do this all the way through school, we'd only have half as many people dropping out at the bottom. Now I'll show you how to do that. It will be in a limited context, but I'll show you how to do that.

We do it by revising the content from feedback that we get by mining the data that is accumulated in the online system in the course of the students using it. We originally gave the new carefully edited problems in cybertutor.mit in 2001. Four percent of the students here just abandoned the problem. Twelve percent requested the solution. Students averaged 1.5 wrong answers per question, and requested 0.75 hints per question. Then we spent additional time improving each problem based on data on student performance – about 10 percent of the time it took us to write it originally. In 2002, we gave that class the improved problems, and look what happened. Only half as many abandoned the problem. A little less than half as many asked for the solution. The number of wrong answers per part went down 40 percent, almost a factor of two. The number of hints per part went up, because we added a few hints where they were clearly needed. And the students spent exactly the same amount of time on each question.

Frankly, I think the students budget their time pretty intelligently. Each student is going to spend a predetermined amount of time on physics. If that means at the end they are going to have to copy a lot, well, they copy a lot. We also know that the copiers don't start their work in a timely fashion. They don't even finish it on time either. But let me get back to this. The next year, Physics Tutor was a commercial product with a vastly improved interface and some editing of the problems. We did not have access to all of these data and could not analyze all the other bars. However, the number of kids who didn't finish went down further. This is an example of the power of analyzing and thinking about the data that you mine out of online systems.

Here is a way to incrementally improve our central course element - the sage on the stage and everybody there with notebooks. The actual result of this educational element is to transfer information from the professor’s notebook to the student’s notebook without passing through the minds of either one. That is how some people have characterized a lecture. They are actually pretty close. I mean, most of you are paying attention— which is not like a typical lecture at MIT, I hate to say. But if on Friday you had a test on this, research shows that you would remember about 10 percent of the new concepts I mentioned. Here’s an example of what kind of product we should be creating to improve the learning. We start with a video, for example, of a lecturer with some demonstrations and whatever else he has, and we have standard “start” and “stop” buttons. Then down on the left here, we have a streaming text of what is being said. Some stuff similar to this already exists in OCW.

Here is what we do differently. We have “Frequently Asked Questions” that are relevant to this lecture, at this point in the lecture. These questions are scrolling by, and students can pick ones of interest. Then we use techniques like those used by Google to re-rank all the “Frequently Asked Questions”. Maybe we have a TA behind this. The students can ask new questions. The students can suggest URLs that helped them understand this. These “Frequently Asked Questions” and the helpful— supposedly— references are shown to other students, and we look at what helps the students understand. Where do they go? How long do they stay there? And if we reconfigure things so the FAQ’s are assessment items, there’s a very important fact that we know that Google does not. That is, “Could the student answer the question when they came back from the supposedly helpful resource?” In this way, we get the system to be more scalable, because every student is indirectly helping every other student. If you are in a country where you have one teacher for every three or four classes, this is the kind of stuff you have to do. Finally, in my suggested system, whatever the student highlights of all of this information, can simply be added to their notes. There is also a course roadmap, because giving students an overview is really one of the hardest things in any course.

I've talked about my zero based thinking here today, and emphasized universal assessment. I’ve talked about improving the content and how we might measure its effectiveness. I want to end by saying that I think we have to start using social network stuff much more to help students educate each other. I just hired a post-doc who had done a thesis using something called Galaxy Zoo. A grad student realized that it would basically take him forever to classify the 20 million galaxies in the Sloan Digital Sky Survey. Well, are these elliptical galaxies? Are they spiral galaxies like the one we live in? Are they barred spiral, with a little bar in the middle? There are about 12 basic kinds of galaxies. Actually, this post doc discovered a new kind of galaxies called green pea galaxies, using this Galaxy Zoo. People participating in Galaxy Zoo get a little bit of training, they come in, they try to classify these things, they have discussion groups, and basically the system ranks how helpful they are to other people, how long they've been there, and how good their answers are. So the people in this network function as ranked peers.

The idea that the students learn from each other is not new. My mother, in 1927, taught in a one-room schoolhouse in Vermont. I once asked, “Mom, how did you ever teach K-8 classes simultaneously?” She said, “You give the kids an assignment. If the second graders can't do it, they go ask the fourth graders.” Duh. That ought to work. Even if you have a class filled up, and no teacher for it, that can work. All you have to do is get someone in there who can implement that idea. They do not have to be the subject expert. The Web and done self-learner can be the subject expert.

There is no reason you cannot have student groups produce the content, use the system to assess its appeal and judge its educational effect, help the student groups improve it, and educate tomorrow’s learners with attractive and effective free materials. From my point of view, that is the future of education in the digital age. Thank you for listening to the end.