Assessment, learning and technology: prospects at the periphery of control

5 September 2007

Transcript of a keynote speech by **Dylan Wiliam, Deputy Director of the Institute of Education**, at the 2007 Association for Learning Technology Conference in Nottingham, England. In the chair, Sharon Waller, Anglia Ruskin University. Slides for this talk are at http://www.alt.ac.uk/docs/altc2007_dylan_wiliam_keynote.pdf [400 kB PDF]. **Slide transitions are indicated in square brackets.**

This text transcript is at http://www.alt.ac.uk/docs/altc2007_dylan_wiliam_keynote_transcript.pdf [75 kB PDF]. An MP3 recording of the talk is at http://www.alt.ac.uk/docs/altc2007_dylan_wiliam_keynote_audio.mp3 [13 MB MP3]. Slides and video of the talk, captured as an Elluminate Live! session, is on the ALT-C 2007 web site at http://www.alt.ac.uk/altc2007.

Good morning everybody. I'm Sharon Waller, co-chair of the 2007. And I'm delighted to introduce our second keynote speaker, Professor Dylan Wiliam, who will be exploring some of the ways in which technology will change how learners are assessed. Professor Wiliam is deputy director of the Institute of Education in the University of London and Professor of Educational Assessment. In a varied career he has taught in inner city schools, directed a large scale testing programme, trained teachers, served a number of roles in university administration, including Dean of a School of Education and pursued several research projects, focusing on supporting teachers to develop their use of assessment in supporting learning. From 2003 to 2006 he was Senior Research Director at the Educational Testing Service, Princeton, New Jersey in the United States. So you won't be surprised to learn that Dylan is a leading expert in the field of assessment, and is often consulted by the media and government for his views on assessment and education in general. Dylan has many publications to his name, perhaps the one that is most often cited for its influence on classroom assessment practice is Inside the Black Box: Raising Standards through Classroom Assessment, which he co-authored with Professor Paul Black. This followed a major review of the research evidence on formative assessment. As you will hear this morning in his consideration of some of the ways in which technology can change how learners are assessed, and how teachers can be supported in facilitating that process, Professor Wiliam continues to influence policy and practice by pushing the boundaries of our understanding of how best to design assessment in order to improve learning, rather than just measure it. Please join me in welcoming Professor Wiliam to ALT-C 2007.

[applause]

[2] The emphasis in my talk today is going to be very much on the learning, rather than the technology. And I hope to convince you about why that's a good idea. I'm going to start off with some precepts about learning and about teaching. I want to point out that what we really need in classrooms are what Lee Shulman calls "pedagogies of engagement" and "pedagogies of contingency". I then want to turn, in closing the talk, to the role of technology. And I'm going to suggest that the role of

technology lies in supporting, rather than replacing teachers, and I want to suggest to you that the really important and exciting development for technology in supporting learning is in something called classroom aggregation technologies.

First of all, why do we need to raise achievement? Because it actually matters. [3] People with longer education live longer, they have more money, and they are healthier. Society benefits by having lower criminal justice costs, lower healthcare costs, and the economy grows faster, the more educated your population is; it's just as simple as that. So where's the answer? [4] Well this morning the Conservatives announced small high schools as their platform priority. I used to live in New Jersey and our state capital, Trenton, was part of this new small high schools buzz. They took a three thousand student high school, divided it up into six five hundred student high schools in the same building. [laughter] And they wonder why nothing changed. Chicago's got a city-wide small high schools initiative – it didn't work. It improved attitudes – teachers got on with kids better, but they didn't learn any more. Governments like things they can do easily, so they can change curricula, they replace textbooks, they go for charter schools or vouchers or academies or trusts. And of course, technology. Technology has been about to revolutionise classrooms for about thirty years. As Heinz Wolff once said "The future is further away than you think." The latest one of course is interactive whiteboards. Charles Clark, when Secretary of State, decided every school in London should have one, and we were fortunate enough to have the resources to evaluate that properly. And what we found was there was absolutely no evidence of impact on student achievement. Actually, what we found was there was evidence of *no* impact on student achievement. There were as many schools where putting in whiteboards had made things worse, as there were where they made things better.

The problem is that we're looking in the wrong place for the answer. We've had three generations of school effectiveness research. [5] The first one was raw results. Some schools get good results; some schools get bad results. "Ooh, that must mean that some schools are better than other schools." So the conclusion was schools made a difference. Then people said, "Hang on a minute. The schools where the kids are getting all the good results, they're the ones in the posh areas." So people said "Ah, right, okay. So we'll control for social class." And what did they find? Social class actually accounts for most of the variation. So the conclusion from this was that schools don't make a difference, it's all to do with poverty. And then people said "Hang on. Why don't we actually look at what the school contributes?" Let's look at how much the kids knew when they started at that school, and how much they knew when they finished—the value added approach. And what that has shown is that actually it doesn't matter very much which school you go to, but it matters very much which teachers you get in that school. [4] The variability at teacher level is about four times the variability at school level. If you get one of the best teachers, you will learn in six months what an average teacher will take a year to teach you. If you get one of the worst teachers, that same learning will take you two years. There's a four-fold difference in the speed of learning created by the most and the least effective teachers. And it's not class size, it's not between class grouping, it's not within class grouping – it's the quality of the teacher. So we have in classic economic terms a labour force issue, with two solutions. We can either sack all the teachers we've got and start again like Ronald Reagan tried with the air traffic controllers. Nice idea. Unfortunately there aren't any better teachers out there who are deterred by burdensome certification

requirements. And actually, new teachers are actually pretty bad. You don't really learn to teach at all well until you're six or seven years into the profession. And some recent data from Australia shows that the amount of value added by teachers actually carries on increasing for about twenty years. Basically almost all teachers are almost useless when you start. [laughter] And you're halfway decent by the time you finish. There's nothing harder than teaching. And you're hardly ever successful. You show me a teacher who's satisfied with what they're doing, and I will show you a teacher with low expectations. Our constant experience is of failure, but by learning, by practising we can actually get better. So the only way to improve learning on any kind of scale is to improve the effectiveness of the teachers we've already got—what my colleague, Marnie Thompson, calls the "Love the one you're with" strategy. How do we do it? And what is the role of technology in that? [7]

[8] Well, in the past I've talked about the difference between quality control and quality assurance. [9] And you know that basically quality control is a kind of "bolt-on" thing, where you actually inspect things coming off the end of the production line. You look at the things and if they're any good you let them go. And if they're bad, you send them through the production process again. So you "inspect-in" quality and everybody thinks this is very bad. Quality assurance is good, because you build quality into the process, so there's no need for inspection at the end of the process, because you design quality into the manufacturing process and quality is designed in, and therefore that's good. [10] For some processes quality assurance is more efficient than quality control, e.g. automobile manufacture. That's why Toyota is the most efficient auto manufacturer in the world. They built quality into their production. But nobody has yet managed to do that for silicon chips, as far as I understand. The most efficient way to make silicone chips is actually to make lots of them, and test them, and throw away the ones that are useless. So the crucial trade off in whether you go for quality control or quality assurance are to do with testability, complexity and predictability. And the question is: where does learning fit? Now Brenda Denvir—for a PhD thesis back in 1986—produced an extraordinarily detailed map of young children's acquisition of number. So each of these little blobs [11] is a skill. So for example what she showed was that there are almost no kids who could do subtraction without being able to count backwards by one. So counting backwards by one is a prerequisite skill for subtraction. She mapped this very accurately and what they did was, they looked at programs designed to help children learn. Now if you look at this map here, here are the prerequisite skills. These arrows are dependencies, and obviously the kid knows this, the kid knows this. They know this, they know this, but they don't know this. So this is obviously a target for teaching. The teacher then designed a program of teaching specifically to address this skill for this child, one to one. What happened? The child learned those things up there. That's what got learned. We cannot—anybody who's been in a classroom more than a nanosecond actually knows this—we cannot predict what it is that children will learn as a result of our teaching. So we cannot have quality assurance in learning. We have to have quality control. We have to keep on checking on what it is the kids have actually learned, because we cannot predict it. You cannot have perfect teaching.

There was a craze in America a few years ago for perfect teaching, where they would give these teachers scripts, you know, designed by experts on how to teach really well. And they were really scripts, there were things like "Now, walk around the classroom". [laughter] And the point is, they were useless because classrooms are

effectively chaotic places. Actually, even well behaved classrooms are chaotic places, in that the difference between one course of action and another course of action is so small that it's effectively only described well by chaos theory. So you cannot prejudge the complexity of the situations that teachers will face. So what gets learned? Well here's another slide which shows some items from this Third International Maths and Science Study [12]. They're both about which fraction is the largest or smallest. The first item, eighty-eight percent get it right; the second item, forty-six percent get it right. And it's not the fact that the numbers are bigger in that second question. It is the fact that a lot of the kids have the naïve strategy the biggest bottom makes the smallest fraction. They got the right answer in the first question, but not in the second.

Which fraction is the smallest?
Look for the biggest bottom
Find six
Choose (a)
Correct.

Which fraction is the largest?
Look for the smallest bottom
Find four
Choose (b)
Incorrect.

So what gets learned is actually very, very difficult to predict.

This [13] shows how slow learning is. We tested some kids over a five year period, asking them basical mental arithmetic tasks—they could actually make some notes—what is 860 plus 570? At age six and a half, fifteen percent of the kids can do it. At age eleven and a half, about ninety percent of the kids can do it. And I think most people would be surprised by how flat that line is. That every year, only about fifteen percent of the kids are getting this—fifteen percent! So in a class of thirty, six kids are getting it this year – one every two months. The SESM projects [14] found that typically in teaching one third new the content at the beginning, one third didn't know it at the end, so only one third learned the content, and half of these had forgotten the content six weeks later. [laughter] But you know this. Perhaps more surprisingly, some did better on the delayed post-test than on the immediate post test. Some kids didn't know it at the end of the teaching, but they did know it six weeks later.

The important thing is that what gets learned as a result of a particular sequence of instruction activities is impossible to predict, but student errors are not random. Those are the two most important insights in twentieth century psychology [15]. And most of our pedagogy is designed around the idea that students' errors *are* random. When kids don't get stuff, what do teachers do? They do it again, but slower and louder. It's a model based on association, and the idea is, actually. quite, quite respectable psychology. The idea is that learning is a process of forming links between stimuli and responses and therefore learning is assembling these chains of stimuli and response. And if they haven't learned something, well the problem is that those links aren't strong enough. So you reinforce, so you rehearse. So repetition is actually the right thing to do if that's what happens when learning takes place. But it's not for most of the kinds of learning that we're interested in. It is actually probably good

model for learning the times tables. But it's not a good model for science learning and maths learning. The conclusion from this is that teaching is interesting because learners are so different, but only possible because they're so similar. And that's why learning is a 'liminal' or 'threshold' process at the boundary between control and chaos. You cannot respect the individuality of every single child, but also you don't have to. The difficulty is learning to cope with—and reducing—that complexity into something that's manageable. It's also why all that research on learning styles is completely fruitless.—loads of stuff on learning styles and students doing VAK inventories—it's all a waste of time. Partly because it's impossible to actually cater for the individual needs, and secondly it's not even a good idea. Can I ask you all to fold your arms? Now do it the other way. Learning in your preferred learning style is like folding your arms the way you like doing it: it's comfortable, it's natural, it feels easy. Learning outside your preferred learning style is like folding your arms the other way, and it feels really weird. But what's interesting is you actually then start to have to think about what is involved in folding your arms. And doing it the way that you don't find comfortable actually gives you more insight into what is involved in folding your arms, than doing it the way you like. So what's really important is kids need a balance between being inside and outside their preferred learning styles. And you don't need to know which kids are in which stage at which time. You just need as a teacher to vary your teaching style.

Now I want to spend a little time talking about learning power environments [16]. Learning power is a concept that Guy Claxton has put forward. The key concept here—the big trap—is that teachers do not create learning. That's true—teachers do not create learning, and yet most teachers behave as if they do. Learners create learning. Teachers create the conditions under which learning can take place. Our schools don't function like that, which is why somebody once joked that schools are places where kids go to watch teachers work. [laughter] And certainly with the intensification of test results, I see teachers working very hard—if the teachers are going home more tired than the kids at the end of the day, the wrong people are doing the work. [laughter] The crucial feature of well-regulated, well-engineered learning environments—and I think it's important way to think about this—is about creation of effective learning environments. It is an engineering process. The key features are they create student engagement and they're well regulated. And I'm going to say a bit more about each of those.

Why engagement? [17] Well, it turns out that intelligence is partly inherited. So what? Actually if you read the media, you'd think that wasn't true, but every single psychologist who knows the data, knows that actually there is an inherited component, and it's not zero. But it's also partly environmental, it's like physical height. Taller parents do have taller kids, but the height the kids eventually reach is based on a whole other range of factors, such as nutrition, and the rest of it. We've always known that environment creates intelligence. What we haven't understood until recently is that intelligence creates environment. It turns out that intelligence becomes a better predictor of people's jobs, the older they get. That's completely counter-intuitive. You'd expect the importance of intelligence to be less and less important as people get older. Actually it becomes more important because people choose for themselves cognitive niches that match their preferred level of functioning. And kids do the same in classrooms. In some classrooms there are kids who are trying to actually answer every single question the teacher is asking. And those kids are

actually getting cleverer—their IQs are going up. Neil Mercer of the Open University has shown that when kids engage in meaningful dialogue in science lessons, their IQs on Raven's progressive matrices—which are purely spatial IQ tests—go up. There are other kids in the same classroom, who are trying to avoid being asked a question. Those kids are foregoing the opportunity to get smarter. So if any teacher is allowing kids to choose whether to participate in the classroom discussion or not, you're actually exacerbating the achievement gap. That's why we need pedagogies of engagement, where we create learning environments where there's a high cognitive demand, which are inclusive of all students, and where participation is obligatory. And a good example of that is the work of the Hungarian American psychologist Csikszentmihalyi, who invented the concept of flow [18]. The interesting thing about his work is he completely turned around the research on motivation. Most psychologists up to that point had treated motivation as an input. Some kids have it, some don't. Kids who have it do well, kids who don't have it do badly. But he said actually motivation is an outcome. When you give kids challenging stuff to do which is just at the level of challenge they can cope with, they will be motivated and they will actually get this sense of flow. Whereas if challenge is low, they can become bored. And he documented lots of examples of mountain climbers, ballet dancers, chess players, who talk about getting lost in an activity. Those of you who have been involved in computer programming will know, it's that idea of "I'll be with you in five minutes dear". And three hours later you actually think that it is only five minutes later and it isn't.

So pedagogies of engagement are important, but why pedagogies of continengcy? Well as I said earlier, it's because learning is unpredictable. We've done a good job of actually getting assessments that evaluate institutions, and that describe individuals [19]. But we haven't done a good job of using assessments that actually support learning. And that's why formative assessment is so important. It's because we can't predict the learning, therefore we have to monitor the whole quality of the learning constantly while it's taking place. Now that's just my opinion, but the research says it's actually the most effective improvement you could make to teaching. So beginning with the Gary Natriello's work in 1987 [20], within the last twenty years, these studies are syntheses of, between them, around four thousand research studies. And they find consistent substantial effects. I want to focus on Jeffrey Nyquist's work, which is not very well known because it focused on higher education [21]. But he looked at different kinds of feedback:

- knowledge of results
- knowledge of results plus knowledge of correct results (telling kids what they'd got wrong and what the correct answers were)
- giving them an explanation of some kind.
- giving them specific actions to take for reducing the gap between where they are and where they need to be; and, most sophisticated, what he called
- strong formative assessment, where you actually give them an activity to do to close the gap.

What's interesting is that he found about a hundred and eighty studies [22]: 31 on weaker feedback with an effect size of 0.14 standard deviations and feedback: forty studies, average effect size 0.36. The important thing about this table is the more effectively that principles of effective formative assessment are instantiated, the

bigger the effects. I think what's interesting from the point of view of learning technology, is that most of the learning technologies have actually got stuck in the feedback only move. And that's why the effect is disappointing. You get something like twice the effect when you actually find ways to do activities that close the gap.

Just how effective is this? Well, one of the things that I think we don't do very well in education is actually to do cost benefit analyses [23]. We say "This is having a significant impact on student learning". So, how big an effect and how much did it cost? Class size reduction—reducing class size by thirty percent—actually gives you a twenty percent increase in the speed of learning. But it costs twenty thousand pounds per classroom, per year. If you increase teacher content knowledge by one standard deviation, you get a five percent increase in the rate of learning—actually very small—smaller than most people would think. And nobody knows how much it costs, because nobody's managed to do that yet. But if you get teachers doing formative assessment in their classrooms, you get a seventy-five percent increase in the speed of learning and it costs about two thousand pounds per classroom. So that's why I'm advocating formative assessment.

Now can technology help? Rather than, "Technology is the answer. Now what's the question?", the search for me has been around what kinds of roles can technology play in helping teachers do effective formative assessment in the future? Because that's the place where we're going to get the really big impacts on student achievement. I'm going to talk about three generations of pedagogy [24]. The first generation of pedagogy is traditional pedagogy, which is the kind of chalk and talk. You've negligible contingency; I just say stuff to you, and I hope you get some of it. And I can actually polish my presentation and maybe get it better. But there's no feedback at all. The second generation is "All-student response systems". So as I go, I collect information on everybody. And the contingency, the degree of contingency depends entirely on the teacher's skill. What I'm going to argue is that the role of technology in improving learning is primarily in what I call third generation pedagogies. Where we have automated aggregation technologies, which actually take the responses of different students and do some smart things with those things. And give the teacher advice about what are the sensible next steps. The really brilliant teachers are doing this already. But most teachers can't do it. And so the challenge of third generation pedagogy is to have the contingencies of teaching—that what you do when you know that the teaching didn't work quite the way you intended—that is supported by technology.

There is a paradigm evolving in America called "evidence centred design". Basically you design assessments starting from what it is you want them to do. That doesn't sound very radical, but for assessment it is. And Almond, Steinberg and Mislevy have invented what they call a "four process architecture" for assessments [25]. You have the selection of tasks, the presentation of tasks to the assessee. You identify evidence arising from their performance and you find ways of accumulating the evidence. And I'm going to say a bit about each of those in turn.

So this is about questioning [27]. Here's a question: look at the following sequence. Which is the best rule to describe this sequence? Well the correct answer is all of them, because depending what n is, any of these could be right. I don't learn anything from your thinking by just knowing which one you choose. I have to have some sort

of “Well why did you choose that?” So I have to get some reasoning. Compare it to this item here [28]. “In which of these right angled triangles is a squared plus b squared equal to c squared?” [answers from audience] No it’s not all of them. No. It’s B *and* D. Now if I’d given you lettered cards with A, B, C, D, E and F on them, you have to hold up the correct answer, that would be an all-student response system. There would be nowhere for you to hide, because I’d say “You haven’t given me a choice yet”. And with smaller audiences than this I really do that. And I’m about to do it again for you later on. But the point is this—B and D are the correct answers. So if you hold up B and D, then you’re correct. And if you get anything else you’re wrong. But you might be wrong in an interesting way. You might just hold up B. or you might say all of them. But what I’m saying is by having crafted this question in a smart way, just from knowing what you chose, I get very, very good information. If I’m teaching, I can do a quick check and if everybody gives me B and D, I move on. If everybody gets it wrong, I do it again slower and louder like most teachers do. But if half of you get it right and half of you get it wrong, I can then say “Well you thought it was B and you thought it was D. Why?” And you can have a good discussion. But the point is, the design of the question allows you to make those very strong inferences, and your chances of getting this question right by guessing are incredibly small, because there are six possibilities and the solution space is two to the power of six. So the chances of you getting the right combination by guesswork is one in sixty-four, not 20%, as it is in typical multiple choice for example. So the right tasks are important. Now what is interesting is you can’t use this item with the clickers that are proliferating over higher education because most of the systems available only allow one correct answer. And so therefore you’ve got this problem of kids getting it right by guessing. Whereas well designed questions with multiple correct answers give you a very, very small solution set compared with the whole space and give you a really strong warrant. Here’s an example from science [29]. Ice cubes are added to a glass of water. What happens to the level of water as the ice cubes melt? All the answers are correct. A can be true if there’s evaporation, B can be true if you are a good physics teacher. C is a good answer if the ice cubes weren’t floating but were piled up like a scotch on the rocks. And D is actually the correct answer because I didn’t tell you what temperature the water was. It turns out of course that B, the physics teacher’s answer is actually not correct. Because what happens when ice melts is it cools down the water. So the water shrinks, unless it’s between zero and four degrees Celsius, in which case it expands. [laughter]. This is a great question to have a good discussion about. But there’s no point asking this question unless you’ve got the time to hear people’s answers and get their arguments. In comparison to this one [30]. The ball sitting on a table is not moving. It’s not moving because:

A. No forces are pushing or pulling on the ball. That’s the common misconception.

B. Gravity is pulling down, but the table is in the way. Can’t see anything wrong with that, can you? Gravity pulling down? Yes. Table in the way? Yes.

C. The table pushes up with the same force that gravity pulls down. That’s obviously what the science teachers are looking for.

D. Gravity is holding it on to the table. Hm, that looks pretty good too.

E. There's a force inside the ball keeping it from rolling off the table. That's not correct obviously. But it's an interesting misconception. It comes from children thinking about inertia as a force, rather than a property of matter.

But the interesting thing about this question is this is a great question for checking on students understanding of physics, because B and D are correct, but not physics. And if all the class give the answer C, then you know they've got the point you were trying to get over about the opposition of forces in equilibrium. B and D are actually correct, but they're not physics. So again, this is a very, very high powered question and Mark Wilson and Karen Draney at the University of California Berkeley have got lots of items like these which are incredibly powerful. And kids almost never get them right for the wrong reason. So that if you actually get the right answer from kids, you know you can move on quickly. But they're incredibly hard to come up with, these items. I quite like this question [31]. "What can we do to preserve the ozone layer?" And you read through:

A. Reduce the amount of carbon dioxide in the air
B. Reduce the greenhouse effect
C. Stop cutting down the rainforest

And then "Properly dispose of air conditioners and fridges." Doesn't that look like an item where the item writer ran out of ideas for the last option and actually thought "I'll put something really stupid in there to see if they're awake." Unfortunately it's the correct answer. It's the only correct answer. Because the others are all about the greenhouse effect, not the depletion of the ozone layer, which was caused by the proliferation of chloro-fluorocarbons, which were caused by improper disposal of air conditioners and fridges.

So these good questions can take a little while to come up. In English, you know you've got the traditional English literature question, Macbeth – mad or bad [32]. Great question, but you need to discuss it.

Where is the verb in this sentence? [33]. A very good all student response question. This one [34] is even more sophisticated. "Which of these is the best thesis statement? Now this is only relevant to a particular genre of writing known as "persuasive writing" in the USA. C is actually the best thesis statement. They're all credible thesis statements, but there are some ones that are not as good as C. And so the important thing is that—the teacher knows that if kids choose C and reject D, which is also a thesis, as is E (but it's not a thesis within the genre of persuasive writing)—the fact that the plausible distractors are so good is what makes it a powerful item.

But these items are very hard to come up with. We call these hinge questions. They're questions that are based on an important concept that is critical for students to understand [35]. And when we use them with teachers, we say to them they must be able to collect and interpret the responses from all students in thirty seconds. So you can't get kids to explain their answers. Teachers always say to me "Oh I'd get every child to explain their answer." But they never do, because by the time you've heard from the twenty-third child, the rest of the class is losing the will to live. And so they never do hear from every child.

So what we're saying is that first of all if we're going to have technology helping learning, the first thing—and what technology cannot help us with, and what the clickers can't help us with—is questions that are worth asking. And that's a skill and a craft that we're actually only beginning to get to grips with. You can use these kind of questions for very, very low order things. For example, instead of giving the kid a test on figurative language, you just give them cards with A, B, C, … H on them and you just read out the statements on the right [36] and just say "He was a bull in a china shop." Which is that? And you hope the kids say it's a metaphor because the word "like" is actually missing. And "May I have a drop of water?" It's actually none of these, it's *litotes*. But the point is, you can actually run through these very quickly. And of course, some of these have two correct answers like "the sweetly smiling sunshine" is personification *and* alliteration. So these good questions can actually help teachers make instructional decisions in real time. And it's these kinds of decisions, these kinds of adjustments to student learning, at a whole class level, is what the research has shown makes the biggest difference in creating both pedagogies of engagement and pedagogies of contingency. Because when you actually require a response from every single kid, it has nowhere to hide in the classroom. So everybody has to be engaged, and the teacher is constantly adjusting their teaching. Maybe five percent of teachers can do this currently. Now what I'm suggesting is that we need to move towards more sophisticated methods of evidence identification [37, 38]. I mean currently the great teachers do this with dry erase boards. "Everybody hold up an answer." "Give me a fraction between one sixth and one seventh." "Write it down." "One over six and a half?" Interesting answer; shows me that they're thinking.

But we need to explore the use of technology in order to capture that information, so that we can actually begin to do something smart with it. So we've got classroom clickers, we've got the traditional keyboards, wired and wireless, and anoto pen. Here's a classroom with a set of the low tech version [39] because you have ABCDE cards on a string attached to the chair. And the teacher just says "Okay. Reach for your cards. Give me an answer." So it's always there.

I don't know if you've come across the Anoto pen [40]. The Anoto pen is very smart, because it knows where it is. So you can say for example, if you care about such things, give the kids a map of Britain, just put a cross where Manchester is. And the kids are doing it at their desk, on a piece of paper. Put a cross there, and the teacher can actually see where all the crosses are. That is the beginning of classroom aggregation technology, because it's when the teachers can begin to aggregate the information from different students—that's the evidence identification. The Palm wireless keyboard [41] is being used in lots of classrooms in America and the classroom clickers—as I said, the next generation will presumably have a facility for multiple correct responses [42]. Discourse [43]: does anyone know a software package called Discourse? This is a very interesting example, because you have a screen, the kid has a screen like this, where there's a question and the kid has to type a response on the screen. And then the teacher has this screen here where they can actually see the kids' responses as they're writing. So you can actually listen in on child number thirteen and say "You haven't written anything in a little while". But the other thing is you can actually then project one child's response to the whole class, either anonymously or with attribution, and use that as a focal point for discussion.

With multiple choice questions, you can have them scored automatically by matching with the key. So this is a good example currently of an aggregation technology, but it doesn't allow evidence synthesis except for multiple choice questions. Now some of the really exciting stuff is happening at the place I used to work—ETS—where we have evidence identification software for non-multiple choice answers [44]. There's a package called e-rater, which does automated essay scoring. And it now scores essays more accurately than humans. Scary thought, but actually it's not so scary once you realise how bad humans are at marking the stuff. But basically, and in many high stakes exams now there's one human marker and one automated marker, and if there's a difference it goes to a third human—a second human for third marking. But what's interesting is how it does this because actually what most people pay attention to are very broad surface level features like grammar usage and mechanics, spelling, style and organisation. Like you know, is the last paragraph saying "in conclusion" or "finally, to sum up". And it's incredible that just a package that just looks at those kinds of things actually captures almost all of what teachers look at. It doesn't look at meaning at all. At the other end of the spectrum, there's a product called c-rater which is actually a paraphrase analyser. And what it does is takes a short answer question to a question like "What are the important principles in photosynthesis?" And it looks at what kids have written, and it tries paraphrases of those and sese whether it matches the list of right answers they've got. And these are being used in high stake examinations, in for example Indiana. But they're quite limited, because you have to choose your questions carefully. And in the Indiana National Assessments they also use a package called m-rater, which is marking graphs and equations, in an automated way. The problem is that all these technologies are really good for summative assessment. They're not good for formative assessment, because they only help you get a unidimensional answer.

So what we have here [45] is a chart showing what I think is the current situation. Multiple choice technology is useful when you have highly structured evidence. You know it's either going to be an A, B, C, D or an E. And we actually manage to do a lot of work with highly structured evidence. So this (horizontal) dimension is whether the evidence is structured. And this (vertical) is the degree of teacher mediation necessary for the aggregation. Now the ABCD cards are highly structured and therefore you could actually have teacher aggregation. This is, teacher looks at all the ABCD cards. But because the evidence is highly structured, clickers do the aggregations pretty well already. The big goal is to get something happening up this top right hand corner, because what we need is automated analysis and synthesis of unstructured information. And the reason that's so important for formative purposes is because currently we're only very good on accumulating evidence for unidimensional student models. What we're saying is "Let's use all the evidence to put the students in rank order. He's good, he's not so good, he's in between." And what we do currently is build these unidimensional student models [46]. They're useful for summative purposes, but they're almost useless for formative purposes because all you know is "This kid needs to be better." Well it's like telling a bad comedian he needs to be funnier. It's not helpful. It's true, but it's not helpful. If we're going to get serious about formative assessment, using technology, we have to develop multidimensional student models. And this is where the evidence centred design becomes very useful, because we need Bayesian inference networks. So what we would do is build a proficiency model which is actually what proficient performance looks like. We build a task model which is how the task that we're setting relates to the notion of

proficiency. The evidence model defines how we go from the outputs that the student produces towards evidence of achievement, and then what we do is, we use Bayesian inference to update a student model. So the current cutting edge in this area is trying to build student representations of knowledge, trying to build models of what it looks like to be expert in this area; trying then to develop tasks that elicit that evidence, and how in real time you might capture evidence of student achievement to update those models. Because then you can start using the hardware and the software—rather than the teacher's bandwidth—to parse the information, so you can actually at some point either say "The whole class needs to do this" or divide the kids into the following sub-groups [47]. So I can see, at some point software which actually just prints out at the end of a lesson a seating plan for next lesson, where the kids say, these four kids need to work together, these five kids need to work together. And it's all based, not on a multiple choice test, but on a constructed response that the children made that was collected automatically, interpreted and then obviously some individualisation after that. But that's the hope. That's the vision, the possibility that technology I think holds out in really supporting learning. It is teacher mediated, teacher supported classroom aggregation technology. So to summarise [48], I've argued that raising achievement is important. To do so we have to change what happens in classrooms, and we have to work with rather than replace teachers. Specifically, the research evidence shows you get more improvement in student learning when you change teacher pedagogy than when you change subject matter knowledge. I've argued for the importance of pedagogies of engagement and pedagogies of contingency. And I think the role of technology comes in helping us move from single student response systems towards all student response systems, where we're collecting information from all students in real time, and updating a student model constantly. But it's not working automatically, because I don't think the technology is there. A footnote – I don't know whether you've come across a product called the cognitive algebra computer, developed at Carnegie Melon University. It's the only piece of educational technology that has been shown to make a real difference in a wide range of settings to student achievement. You get effect sizes of 0.4 to 0.7 standard deviations. It took twenty years to develop, and it's good for two out of the four or five hours per week that kids get on algebra in grade nine in America. For twenty years, just to get good results for two hours a week for one year in school. And so while that kind of stuff may long term have a future, I think that if we're serious about using technology to support learning, in the short to medium term it is going to require a focus on classroom aggregation technology. Thank you.

[applause]

SW: Thanks very much. We started a bit late, so I'm hoping we're going to have ten minutes for questions. We've got some roving mikes, so if people have any questions they'd like to put to Dylan, please raise your hand so we can see you. We've got somebody down here at the front. While we're waiting for the microphone to come down here, we'll take one of the questions from our remote participants, if that's okay Dylan. We have someone from Napier University in Edinburgh. He asks "Are there differences between motivation to learn and motivation to perform, in other words should we be assessing performance or learning, or both?"

DW: There's quite an extensive literature on this. People like Carol Dweck have looked at performance orientation versus mastery orientation. And there's no doubt

that performance orientation is actually in the long term harmful. So what we need to do is focus students on learning, rather than getting good grades which is why grades have been so deleterious to student learning. So there *is* a big difference and we need to focus on getting students to understand that the really important thing is learning, not getting a particular score or a grade.

SW: Thank you. Gentleman down here, if you'd like to say who you are.

Q1: Thanks for the awesome presentation from your side Professor Wiliam, and my name is Kanishka Bedi and I'm from Universitas 21 Global. We are a purely online institution and we run courses which are accessible to students world-wide. And in our assessment systems, we try to incorporate authenticity as well as possible, because of our programs, management programs in which there are more than one correct answers and these are equally correct answers. We are at a loss to understand exactly how to use technology, because we cannot use objective type questions. At the same time we understand that in order to maintain the authentic approach in our case studies and the kind of assessments we do, we are not able to apply that kind of role technology can play. So what is your suggestion, in what terms technology can be used in this kind of scenarios? Thank you.

DW: First of all the problem with multiple choice is that it's very difficult but not impossible to assess higher order thinking with multiple choice questions. The problem with authentic assessments, particularly the case study approach with a typical student, is that you tend to only assess a small number of cases. So although the reliability, if you mark it again with somebody else doing the marking, may be very high, it turns out the reliability for the student is actually quite poor, because what you're really measuring is "Did they get lucky this time?" Was that a case study they had revised for as opposed to one they'd actually forgotten three months ago? So the reason that multiple choice tests come into their own is when you need to make sure that you're actually checking lots of knowledge in different places. And I would say that any system that is purely predicated on one kind of assessment, is almost bound to be less valid in terms of the inferences that it will support than an assessment that has some bits where knowledge actually is important. So for example, if I was doing this in America, and I was doing accountancy assessment, I would want to know that the people I was certificating knew what the Sarbanes Oxley Act said. And I would do that with multiple choice questions. But I wouldn't rely *only* on multiple choice questions. So it's the diversity of methods of assessment that allow you to get into different kinds of inferences. And that's what makes the assessment more valid.

SW: Thank you. Any more questions from the audience? We have one down here. Yeah, and one up at the top I think. While we're waiting, we have a question from Steve, the TLC group, he asks you if you could explain a bit more clearly what classroom aggregation technologies are.

DW: It's a good bit of feedback. I didn't do a very good job of explaining it first time. [laughter] —the idea of classroom aggregation technology I mean. The human brain inside the head of the best teachers is the best piece of classroom aggregation technology we have. What it does is it synthesises all the information you're getting from different students and actually makes it into a whole course of action for that teacher. So it's about getting information from all the students and collecting it,

aggregating it and synthesising it in some smart way to support action. So those are what I mean by classroom aggregation technologies. And so global warming – fact or fiction? Thumbs up if you think it's fact, thumbs down if you think it's fiction. That would be a classroom aggregation technology, because I can actually summarise the whole group, get you to think about this and say "Okay, what do we think about that? Where do we stand?" So it's any way that allows you to collect information systematically. How do most teachers decide whether they can move on? They ask a question, which they haven't planned in advance, they pick on one kid, who already had their hand up. That kid gives the right answer and I say "Good, well done" and on we go. And so classroom aggregation technologies are the antidote to that. It's actually being more systematic about collecting more information from more of the students before you make instructional decisions.

SW: Who's got the mike?

Q2: Tom Franklin, Franklin Consulting. I've been worrying lately that maybe learning is despite, not because of the teaching. And I suppose the easiest way to summarise that is to say the best prediction of a grade in a subject from a student is the grade they've got in another subject. If good teachers were making a big difference, you'd expect to see them getting As in that, despite getting Ds in something else. But we don't tend to see that. So are teachers really making a big contribution? Or are students actually learning despite the teaching?

A: Well that's where the third generation of school effectiveness studies focus on value added, because intelligence, IQ, that thing that does exactly what you said, which is the kids are above average in English are by and large above average on Chemistry and Physics and Maths as well. And so yes, there is that facet. But we can't change that. But in terms of the difference between what those kids knew when they were eleven and what they knew at sixteen or eighteen or twenty-one, it turns out that—and you're putting a lens onto a very small aspect of what makes a difference—but it turns out the teachers do make a huge difference. So yes, the best teacher will give you a C rather than a D. And the worst teacher will get you an E rather than a D. So it's a small difference for that kid, and, but as I said it's a four-fold difference in the speed of learning. Because basically one year's learning is about one grade in GCSE for example.

SW: Okay, thank you. I think we had a question from up there at the top. Wasn't, Dylan, some of our remote participants are having difficulty hearing you, they say can you speak a bit louder and slowly.

DW: Sorry.

Q3: Derek Marston, Higher Education Academy. Where then is personalisation in what you said? You know the emphasis we're finding in things like the national strategy (what was the DfES strategy). There's this aspect of government policy on the personalisation of learning. Where do you lie on this?

DW: It depends what you mean by personalisation. And I know that there's no consensus about personalisation within the former DfES and its successors. So if by personalisation you mean individualisation, then it's a daft idea because it's not

possible and it's not even smart. But if by personalisation you mean creating different learning environments in which different students can come in in different ways, and that the teaching is adaptive, then I'm totally in favour of it. And actually I would argue that formative assessment is all about personalisation. It's about being more responsive to students, but it's avoiding the trap of individualisation, which is impossible to do because we don't actually allow one to one teaching, because we don't actually pay enough for it. And secondly I'm actually not convinced it's a good idea. We have lots of evidence that children and students often learn better from each other than they do from teachers. So for me personalisation is about opening up teaching, it's about making it more responsive and it's about making more of the students more engaged, but it's not about individualisation.

SW: Thank you. I think we've got time for two more questions. A question from George Roberts "How might these ideas transfer to non-classroom based adult community and work place based learning, with a high degree of learner self-direction?"

DW: I think it would be very difficult. Aggregation becomes much less useful when you've got different people coming to do different things. But I would say that sharing case studies, which we can already do very well, is probably optimum in those kind of cases. So I would say that they are lucky because they're working in a domain where the problems that I'm trying to solve aren't really problems. But it's also very inefficient because everybody's unique path may or may not be interesting to other people. So there's a good degree of inefficiency in that. So I would say it's not relevant to the kind of things I've been saying here today.

Q: Okay.

A5: Tatiana Petrovich. Thanks again for very interesting presentation and bringing back focus from technology to learning. And I'll continue with my question in this manner—you concluded that aggregating technology is an answer for formative assessment in the classroom. But I ask you about readiness of pedagogical theory in terms of interpretation of these answers. Especially for these complex and structured tasks that require lots of prerequisites and lots of pre-skills in order to be successful.

A: The issue you raised is definitely a problem. It's reproducing what experts already do. And the experts already can make sense of this stuff. One of the things we tried to do in one project which is currently undergoing a randomised controlled trial in the United States is to give teachers really high powered questions which focus on misconceptions that students might have. And tell them what are the misconceptions that students are likely to have, and tell them what they mean. So it's a way of bringing on board "just in time" teacher subject knowledge. But I think we need to do a lot more work of actually mapping the kinds of responses that students might make and the kinds of responses you might make to those responses. What is really interesting is that in Japan they have a word for this which is *kyozaikenkyu*. They have a word for the teacher's knowledge of the kind of difficulties that students have with this material, and what to do about it. And I think it's very interesting that they actually have a word for it, whereas we just say it's a problem.

Q: Okay. Thank you very much Dylan, that’s been a very inspirational session, and I’m sure it’s given us much food for thought, particularly around the area of personalisation. Okay, and thanks too to Maggie for moderating so well. So can I ask you all to thank Dylan.

[applause]