Applied methods in the education field: Experiment, surprise seeking, theory

Stephen W. Draper
Department of Psychology
University of Glasgow
http://www.psy.gla.ac.uk/~steve/

Abstract

This paper presents a personal view on e-learning research methods which is essentially pragmatic and above all believes that a mixture of methods is essential. The traditional contrast of quantitative and qualitative is discarded in favour of contrast between comparisons (that can settle questions) and open-ended observations (that discover what questions are important). Both are necessary, and combining them in one study leads to a 2-pass method. The relation of theory to each of these is then discussed. These ideas are then applied to a critique of some other styles of e-learning research. Finally, comments are offered on the relationship of these ideas to a framework for qualitative social science research (which of course is not necessarily an appropriate comparison).

Introduction

In this paper I will briefly describe my approach and views on evaluating e-learning cases. As will be seen, the key elements I revolve around are comparisons (e.g. systematic surveys or controlled experiments), seeking out surprises that overturn our expectations, and the importance of theory in understanding those surprises and reducing the number of future ones.

Comparisons

When we want to settle a question, it is hard to beat a direct comparison. This is true in the education literature too. The most convincing paper in any way related to e-learning that I have seen is Hake's comparison of the "interactive engagement" and "traditional" teaching methods for elementary mechanics in first year university and school courses: so called "interactive engagement" vs. "traditional teaching". Data from 62 courses (6,543 students in all) which had all used a standard Mechanics test both before and after the course were used. Essentially double the learning occurred with the "interactive engagement" method in most cases. The study is quantitative in at least two fundamental ways (besides using statistics to further drive home the point). Firstly, whenever we have a lot of things (mechanics courses in this case), counting is a good way to summarise them, rather than the lengthy ennumerative lists such as those you find in the old testament. Secondly the measure used for comparison is a numerical test score.

However the most important feature is not the use of numbers, but of a direct comparison. The debate over methods is usually characterised as "quantitative vs. qualitative" but the important question is actually about comparable observations vs. open-ended ones. Both are important, and discussed respectively in this section and the next. This is not the same issue (as quantitative vs. qualitative): we can have comparisons without numbers, and in fact some of the best experiments are like this, just as in a race what is decisive is the qualitative feature of who finishes first, not the measured time they took which often varies due to other factors e.g. wind. Conversely, an open-ended study might use numbers as when the time, date, and number of students in a learning lab are recorded, or a UFO observer notes the direction and height of the visual object. These are quantitative measures, but because the conditions are not the same as any other case, the questions not the same, the context not the same, these numbers cannot be directly compared.

Another unjustified assertion sometimes made is that qualitative data is richer than quantitative. If I'm interested in how a student learns from my web materials, then it is true that simply counting the web server accesses is impoverished compared to asking them what role the materials played in their learning. However the fundamental reason for using numbers is to add, not to lose, richness: this comes from the fact that every number has a defined relationship to every other number on the same scale, leading to a huge number of pairwise relationships specified implicitly by each new number given. If I say my grandfather was born in the middle of the Victorian age that is qualitative, but if I were to say he was born in 1860 then (although I used fewer syllables to express the fact) from that single number someone with a familiarity with the period can immediately tell that he was born just after the Crimean war, just before the American civil war, after slavery was abolished in the British Empire, but before it was abolished in the USA or Brazil or Cuba, and so on: a far richer communication. The richness consists of the number of relationships and connections conveyed. Quantities can thus do this, although they will also fail to do it if the numbers are not in relationship to other known facts (e.g. a rating scale from 1-100, but not one used elsewhere). Similarly qualitative data can be rich when we are confident we share meanings with the participant, but of little value when we don't e.g. when a tutor writes "needs to be more critical" on an essay when actually the whole issue is that the student doesn't share a meaning for "critical" with the tutor, or a student writes on a course survey "this course is disgraceful" but doesn't say with respect to what, and anonymity means we don't know what that student says about other courses.

But to reiterate: what is most important in considering our research methods is not quantitiative and qualitative, but the choice between comparable and open-ended observations. While here we might call comparisons controlled experiments, in fact this predilection for them is a much older and more widely held attitude in our culture. Law courts and sports competitions are other examples. The point of the Olympic games, today as millennia ago, is to set up a special occasion where as far as possible the only difference between what two athletes do is due to them: for instance they run the same distance, at the same time, in the same place so all other factors such as slope, running surface, sun direction, wind and so on are the same for all. The Greeks invented and used special technology to get a fair start in foot races: an early technology for improving the validity of controlled experiments. The arguments about whether referees' decisions are correct and impartial in football matches are not arguments about how inadequate matches are in deciding the superiority of one team or another, but about whether the already elaborate arrangements are rigourous enough to provide a valid test.

Returning to Hake's paper: such data seem overwhelmingly convincing. It is one thing to complain about whether a test is fair, or that a success in one school might turn out not to "travel well" to other cases when looking at a comparative study of only one or two classes, based on a test the author themselves created. But here we have tests accepted nationally, and not created by the author, and a huge number of cases.

Another example of a comparison that also seems extremely convincing, this time a deliberate and controlled experiment, is that of Rosenthal's demonstration of the effect of teachers' expectations on pupils' development (Rosenthal & Jacobson; 1968). Carried out in several classes in several schools, they administered tests to pupils and then selected a subset at random but spanning the full range of ability, and deceptively told their teachers that this set showed exceptional promise. Followups over the next years showed that this subset substantially outperformed the rest of their class on measures applied by independent raters, even though the only difference was that the teachers expected them to do better. This makes it reasonable to take teacher expectation seriously as a causal factor in learning. You could find case studies where a learner had a supportive teacher with high expectations and then did well, but this is not convincing, not only because we can never know how well they would have done without that teacher, but because we can find other case studies where a learner did well in spite of lack of support and low expectations in those around them. The latter cases will be taken to "show" that learner self-belief is vital, or that nature is stronger than (poor) nurture, or some other moral. Rosenthal's comparative study however shows that teacher expectations matter substantially when all other things are equal, which no case study, no interviews, no analysis could do.

If we want to settle a question, then we need to do comparisons, to construct a fair trial or experiment based on comparable measures. If the observations are not comparable, then we haven't answered or even addressed the question, only raised it.

Surprise seeking

However that is only one kind of contribution to research. Another major issue is whether we have asked, or even imagined, the right question. This is important in every area of science, and above all important in education and e-learning where the newer the context, the more likely it is we haven't even yet discovered the questions, much less the answers.

In 1978 Penzias & Wilson got the Nobel prize for their 1965 work in which they discovered the universe's background microwave radiation, now interpreted as the pervasive "echo" of the Big Bang. However they weren't looking for it. Their wonderful advanced microwave receiver kept picking up more noise than they liked, so they cleaned out the pigeon droppings from inside the horn aerial, and looked for other obvious problems. Eventually they noticed that as long as the aerial pointed up the sky they got the same noise, but it went away if they pointed it at the ground. Finally they convinced themselves and others that the noise really did come from the sky not their equipment, and from everywhere in the sky, not from particular objects such as the Milky Way. This is an example of discovery completely unrelated to prior hypotheses. I always think of it when I read texts that say science is hypothesis testing, or worse, that it is so important to identify and declare your theoretical stance because it limits what you perceive. Penzias & Wilson's work is a striking counterexample to those views.

Unexpected surprises are if anything even more important, at least in my experience, in research on e-learning. Here are three anecdotes: cases where I myself had my ideas changed by experience, by observation, and not by experiment or comparisons set up to settle a question I was already focussed on.

Anecdote 1.
The evaluation team I was working with were invited to look at the introduction of a simulation exercise being introduced to a Biology lab instead of the physical experiments that formed the rest of lab course. The approximate question in our heads was something like "is the computer simulation as good as or better than the other lab exercises based on physical apparatus?" Fortunately we turned up to observe, despite being briefed by the teacher in charge, instead of doing some analysis of marks or collecting questionnaires from students. The teacher was present essentially (in his view) as a manager to deal with any problems as they arose. But what we saw was that as soon as the student group had finished going through the worksheet for the simulation, he moved across and engaged them in Socratic dialogue. To the teacher this was unconscious, taken for granted good practice at least if he had nothing better to do, so he hadn't thought it worth mentioning to us (or in the jargon, careful interviews of the participating teacher did not reveal this). Our observation however immediately showed that this could not be seen as a comparison of e-learning against non-computer learning: clearly the intervention of a skilled teacher was likely to be at least as important for learning as the technology, and it was unstructured observation, not a prior question in our heads, that had revealed this to us.

Anecdote 2:
I was chatting with a colleague in my office about the structured feedback forms used in our first year course for tutors giving written feedback to students on coursework. The question in our heads was about how feedback could maxmise the improvement in students' coursework. My next student appointment turned up and we were keeping her waiting. Finally we turned to her and asked how she herself had found such feedback in the first year course. She paused, then said that the only question she had (in vain) wanted the feedback to answer was whether she was "really" a psychology student or whether she should change course to geography. In those few seconds she showed me our whole view (and that of most of the literature) of feedback to students was wrong and it could never be understood as a technical matter (in a cognitive psychology framework) of improving their skill, but that it involved questions of identity and social integration: a whole area of theory that I had resisted getting involved with. (It also implied that the whole design of our form needed much more radical change.)

Anecdote 3:
Being interested in ideas of reflection, I asked a friend who was keeping a learning diary if I could look through it. This was a true reflective diary, not one produced as a course requirement for some teacher to assess. My friend, also working on educational research, was taking an HND course (in engineering) to give herself personal experience of being a student again. Within a minute it became clear that about a third of the entries were her concerns with what the rest of the class (mostly male, mostly a lot younger than her) thought of her. This at once showed that the theories of learning and teaching I was most involved with had a large gap in: they had nothing to say about this, yet it occupied the learner's mind as a problem and we must expect that it potentially affected her learning.

The purpose of research is to change first the researcher's mind, and then hopefully the minds of their research community. As the stories illustrate, one of the most powerful sources of this is open-ended methods, that put the researcher in the way of encountering a surprise (by personal observation, by listening to a student giving an open-ended response to a question where she felt able to say something that didn't fit the assumptions behind it, by a personal reflective diary which was used to record what the learner actually thought from day to day, not what fits into Schön's or Kolb's theories of reflection). In fact Hake also provides an illustration of this. While the comparative study referred to in the previous section gives strong reasons why others should, rationally, pay attention to and indeed adopt the "interactive engagement" method of teaching, his own conversion, i.e. the reason why he thought it worth while gathering this evidence in the first place, had another more personal origin: "My present concern with undergraduate science education began in the early fall of 1980 when, being assigned by happenstance to teach a physics class for prospective elementary teachers, I gave the first examination. The results showed quite clearly that my brilliant lectures and exciting demonstrations on Newtonian mechanics had passed through the students’ minds leaving no measurable trace. To make matters worse, in a student evaluation given shortly after the exam, some students rated me as among the worst instructors they had ever experienced at our university. Knowing something of the teaching effectiveness of my colleagues, I was severely shaken." (Hake, 1991). This (while involving the quantitative measures of both exam scores and student ratings) is an open-ended case both because he wasn't expecting or seeking this result, because the comparisons implicit in it (with his expectations, with his colleagues' teaching, with teaching similar material to other classes) are ill-defined, and because its main impact was to cause him to be dissatisfied with his teaching approach and to look for another which at that point he hadn't even heard of. Essentially it is a case study which was convincing to him, but couldn't be expected to have the same impact on anyone else, unlike the later comparative data of Hake (1998).

In most of the studies I've done, it is the open-ended methods that have produced the points of most interest to readers of our reports, partly because our poor understanding of e-learning means we have a long way to go to identify the questions, and partly because it is the identification of which issues to worry about that is most useful to practitioners. It is however difficult to say anything precise and methodological about surprise seeking and how to go about it. For instance the literature on testing (e.g. testing iron castings from a factory, testing a new aircraft, ...) is actually a literature on running experiments to test for the presence of the defects that have been seen before. Yet the real but unspoken reason for most testing especially of new designs, and for using humans not machines to do testing, is to notice the unexpected as soon as possible. But if it's truly unexpected, you cannot have designed a test for it. Nevertheless, general considerations show us that not only humans but animals can do it. If, on the contrary, it were true that we can only perceive what we already know is a possibility, then stage magic would be impossible: we simply wouldn't see the illusion. Learning would be impossible: whatever you didn't know about in advance, you couldn't see and so couldn't learn. Animals would avoid predators they had evolved with, but as soon as a predator evolved a change in appearance, it would wipe out its prey. But in fact we know animals and people are excellent at detecting the unexpected (even though they are still better at detecting the expected), and the emotions of surprise and astonishment are the familiar mental signs that accompany this common occurrence. You (and I would agree with you) may be utterly convinced that it is impossible for a 10 ton fluorescent purple dinaosaur to be seen running down the high street, munching up people as it goes: but if it were there, you would see it and could describe it, and would change your views on what was possible.

In e-learning (and in all) research we need the equivalent, and that is the importance of investing effort in open-ended measures. Our methods need a prominent place for being able to notice fluorescent dinosaurs, even though methodology (the study of methods) is almost entirely dumb about explaining why this is so important. Fortunately practice is often better than the methodology textbooks, in engineering as well as in e-learning. For instance in developing their 777 aircraft, Boeing discovered at least two design problems through open-ended aspects of their testing rather than through analysis, even though aeronautics is in many ways well-endowed with theory: the first time they tested the fuselage it failed to hold air pressure, leading to modification of the door seals, and the first time they flew the new engine, it constantly "backfired" (suffered a compressor stall) leading to a considerable aerodynamic redesign.

However while a single open-ended observation can be enough to tell me that my theories are inadequate, it isn't enough to deal with the more common concern of whether that issue is generally important in this context. For that, we need to convert the identified question into a comparative measure, and survey the whole population. In general, my overall method therefore involves two "passes". In the first, the pre-existing questions (e.g. from stakeholders) are trialled with comparable measures (get all participants to answer on the issue) but plenty of open-ended measures are also used to identify new issues that weren't foreseen. In the second pass, new issues that turned up in the open ended responses from the first pass are turned into new comparable measures so that it can be established to what extent these indeed matter across the whole population or set of cases. By themselves, open-ended ("qualitative") methods can at best expand our ideas of what is possible: they cannot change our minds about what is actually the case, particularly not about what factors are of most practical importance in a given context. For that we need comparable (often "quantitative") measures. Conversely, comparable measures in themselves cannot discover what the right hypothesis is to test, or what is the set of important factors limiting learning in a given situation.

An example of applying the 2-pass method

An example of the 2-pass method is partly reported in Draper & Brown (2004). We got evaluation data from a number of uses of EVS (electronic voting systems) used in lectures, some open-ended and some comparable measures. Among others, we asked (open-ended) questions about what the advantages and disadvantages were e.g. "The anonymity allows students to answer without embarrassing themselves" or "Setting up and use of handsets takes up too much time in lectures". In some later studies, we used these items to form systematic questions. We were thus able to report both on what the positive and negative issues were, but also that while all of these items, good and bad, were perceived by at least some students in at least one of the applications, they often did not apply generally to all uses of the equipment. Furthermore the relative importance of each item changed a lot over different cases. Generally speaking, the benefits stayed fairly stable while the disadvantages changed, as one would hope, as we improved our practice in the light of this formative evaluation data. We could thus report both on possible pitfalls, and that fairly stable benefits could be achieved and so reasonably expected by new adopters.

Another example of comparability became important in these studies, especially in the light of the existence of at least some negative features identified by the open-ended methods. This was that asking students to rate how beneficial using EVS was, was not really useful: much better was to ask for their attitude ratings in a relative i.e. comparable way: "What was, for you, the balance of benefit vs. disadvantage from the use of the handsets in your lectures?" Thus we could both get open-ended information some of which was used to immediately improve EVS use, and comparative information useful to potential users wondering whether adoption would be sensible.

This illustrates that evaluation can be designed for different purposes or roles. (N.B. these approaches could be applied to either e-resources or to teaching methods, or both.):

- Formative evaluation: to help improve the design of the e-learning. Open-ended measures are more frequent here: notice a problem, fix it, re-test; but experiments sometimes expose a problem, and comparability is important to estimate priorities and relative importance of different issues.
- Summative evaluation: to help users choose which piece of e-learning to use and for what. Comparability is central here. Unlike the other roles for evaluation, this is aimed at uncommitted outsiders; either to help them choose between available alternatives in (say) e-learning, or else to help them decide whether or not some new approach is worth adopting. "Fair tests" and comparative measures are the most convincing for these, relative to their concrete question of "should I adopt this?".
- Illuminative evaluation: to uncover the important factors latent in a particular situation of use. Open-ended measures are the only ones here.
- Integrative evaluation: to help users make the most of a given piece of e-learning. While the overall aim is mainly formative, what gets changed is more often what teachers or learner do, or a written resource, rather than the e-resources. I.e. this is about overall practical adaptation, rather than either pure design improvement or choosing between products.

As far as I know the terms, though perhaps not the ideas, were introduced as follows: "formative" and "summative" by Scriven (1967) (see also Carroll & Rosson (1995) for their subsequent use in Human Computer Interaction); "illuminative" by Parlett & Hamilton (1972/77/87); "integrative" by Draper et al (1996)

The 2-pass method discussed above is mainly aligned with integrative purposes. That is, my own overall aim in research is usually to improve learning i.e. it is an applied aim, where practical achievement comes first, and understanding or novelty are secondary aims. However it is not uncommon for pure research to be prompted by problems first recognised in applied research, and in any case pure research also needs open ended methods as much as comparability.

Theory

Is theory of any use, then? The first place it is useful is in diagnosis or interpretation of symptoms or phenomena that may be first noticed through open-ended methods. In some cases the symptoms suggest both cause and remedy directly: for instance if you observe the learners are all weak with hunger and cannot concentrate, then you understand without anything further that they need, not a change to the e-learning resources, but feeding, and you consider perhaps a school meals programme. Often though, theory may help: just as you need some medical training to tell whether the same symptom of a rash is due to the heat or to meningitis, so also if you see the symptoms of lack of work in students of first year computer programming students, theory may help you decide between diagnosing a lack of motivation (the preferred perception of most staff) and a lack of helpful explanation when they get stuck (no point in spending hours on exercises when you just don't understand them and so make no progress). Once alerted by surprise detection, then theory can often quickly catch up and guide a fuller analysis and solution. Again, this is true in engineering as well as e-learning: for instance London's Millennium bridge exhibited an unexpected swaying wholly unpredicted by theory, yet which could be easily understood as well as fixed in retrospect (leading to some textbook rewriting as well). (Of course everyone saw and noticed the swaying: no engineering training was needed.)

However theory also helps better symptom detection. While most uses of health services are driven by a patient noticing an unexpected symptom and then seeking advice, part of the point behind screening programmes and regular checkups of healthy people is to allow theory to drive detection of problems that are not apparent to untrained open-ended observations. Thus theory is in fact an alternative to open-ended observation in detecting surprises, hopefully before too many resources have been wasted, by guiding us to make certain observations in an e-learning study. Nevertheless, we have to deal with the fact of the primitive and fragmented state of theory in education. This point is illustrated by Greg Kearsley's (2006) website. He lists (giving a short description of each) over 50 different theories of learning, without pretending to be exhaustive. The particularly disappointing feature of this is that most such published "theories" make almost no attempt to relate themselves to other work in the field.

At this point we should note the fundamentally contrasting roles of theory in pure and applied work. In pure research, the aim is a truth independent of others, and universal: not just true on particular occasions. Controlled experiments are generally designed to cancel out or exclude other influences to focus on one factor alone. The most grandiose examples come from physics that has gone a long way towards demonstrating that its theories, e.g. gravitation, are true and apply from when time began into the uttermost future, and true here in the room you are now and in the furthest reaches of the universe. However that doesn't mean it is always relevant. Gravity has negligible influence in elementary particle interactions, or on the life of bacteria, because other forces are much stronger in those contexts; and a mouse that threw itself off a skyscraper would probably scamper away, while a horse that did the same would be so utterly smashed it might not even be easy to recognise its species. Gravity is always true and always present, but how important it is varies from utterly negligble to all important, even here on earth. For practical work, whether building a bridge or an e-learning course, what matters isn't a single factor that is always true in theory, but identifying all but only the factors which have a large enough effect to make an important difference in this context.

For example, consider the case of trying to reduce the student dropout rate in a given institution. Theory says (Tinto, 1975) that dropout depends on the degree of social and academic integration, and no-one has got far trying to disprove this. But in the particular institution you are looking at, does that mean you should put on reading parties (group cohesion within a discipline group), put on a peer mentoring scheme (provide social introductions not strongly related to disciplines, that may also lay a foundation for later personal crisis support), train tutors to be more approachable (improve relationships with staff rather than with peers), or provide copious feedback on academic work because a strong sense of succeeding at learning might be more important in identifying with the subject than human relationships are? Should students be offered a lot of choice in courses, or no choice at all because the former means they will have no coherent learning community, while the latter means they work with a fixed cohort throughout their university career? (The average higher education dropout rate in the USA has always been very much greater than in the UK, possibly for this reason (Tinto, 1982).)

In applied fields like education, establishing a theory even with excellent supporting evidence is still a long way from making a contribution to any particular context: we have to test its importance (not whether it makes a difference but whether it makes a big difference compared to other factors), and then to re-test its importance in each case. Progress without theory is possible — in fact most fields of study begin that way. Still, theory, even partial, incomplete theories, amplify our efforts (nothing so practical as a good theory). It is also attractive. Theory appeals to our emotions (that wonderful megalomanic feeling of understanding things), but respect for reality (i.e. data) is more important.

E-learning research

Up to here, I've presented my viewpoint on what is good: the first landscape. In going beyond that and relating it to others' work, the next step is to consider how other work looks from this position and in particular what is bad (the second landscape).

The most prominent research strategy in e-learning is essentially the engineering one of construction. When you build a new bridge, it largely speaks for itself: anyone can travel across it. Similarly thousands of people travel through the channel tunnel each day: they go in one side, come out the other, know personally that it really was built. When Americans visited the moon, the most important "result" was just that fact which anyone can grasp without theory or careful measures: they just went and returned, bringing with them souvenir rocks and pictures. Similarly for the invention of TV or the internet: it constitutes an existence proof. The engineering method could be seen as a variety of case study. It delivers first an existence proof (this kind of artifact is possible and can be built, and the proof is that it has been built); and secondly it provides supportive evidence for the accompanying claims by the engineer about what was important in the design process. It works best when there is no real dispute about the success: that is what makes the case worth studying, yielding a set of features that may have been important contributing causes. In education, still more in e-learning, what is constructed is a learning environment (both materials and the process teachers and learners go through).

The trouble with this research strategy for e-learning work is that often there is no self-evident reason to regard it as a success: the characteristic weakness here is of showing its effectiveness. When you build a learning environment, anyone can look at it, but they don't and in fact can't really learn in it for course credit. So in e-learning the existence proof isn't complete until learning has been demonstrated; and unlike a bridge, the designer cannot be a user of it because they probably already know the material and in any case are not a student on a course requiring it to be learned. A large part of the literature consists in fact of reports of the "look what we did" variety, often without demonstrating its use for learning (like building a big vehicle bridge but only mentioning how pedestrians crossed it). A minimum requirement would seem to be its use in a real (for credit) course.

When you have at least shown its use, you still need to demonstrate that learning occurred since this is not automatically visible (unlike walking across a bridge, going to the moon, etc.).

In fact really to have shown that any effect due to the technology has occurred at all (never mind isolating its important causal features), a trial that is comparative in some way has to be done because we know that people also learn without any help. That is why in drug trials, a control or placebo group is generally required, because whether it is flu, a broken leg, or mental depression, a large proportion of patients get better without any treatment: so the self-evident "getting better" is not itself enough. In this way education is like medicine, but unlike going to the moon, or picking up TV at a distance, or crossing a river where we know that without the artifact nothing happens and an explicit comparison is unnecessary.

Comparability is often the most important issue behind other reported measures. Even with the relatively weak measure of attitudes, reporting that 90% agreed or strongly agreed with the statement "this e-learning resource is useful" is much less informative than reporting on agreement with "this resource is more useful than the textbook". Thus Hake (1991) was devastated by his teaching ratings because they were comparable (and much worse) than other courses the students had taken. This applies even to "technical" measures like weblogs. Saying 90% of the students visited the course website really doesn't tell us much, but it would if it was compared with visiting figures for other course web sites, or attendance at other resources ("90% of students attended at least one lecture in this course", which doesn't sound too good). However to make these figures informative, there has to be choice (of learning resource), and the figures of comparative use given: but this would embarrass many e-learning cases where students are not given any alternatives. Voluntary choice of a resource is quite a powerful indicator, but use of a monopoly supplier tells us little. Of course ratings of utility rather than only use would be even more interesting, but again only if they too were comparable (Brown et al., 1996).

Another annoying weakness could be classed as poor analysis or reflection, but at bottom is again an avoidance of comparison. It is the way so much "modern" educational technology refuses to compare itself with printed textbooks: a technology which although much older in origin, has been continually updating itself. A modern textbook has a contents (for structured access using the conceptual structure the authors designed), an index (for rapid access using the readers' concepts or at least keywords, even when these cut right across the authors' perspective). They can be used not just for the middle of a course, but as a quick preview before, a reference work afterwards, a revision tool, a self-test facility (since today most include self-assessment questions). They can be picked up from a stall and scanned in a minute or two by a teacher considering adoption, or a student assessing their relevance. Very, very few e-learning resources can support all these multiple, but common and long-standing, different user tasks or applications. Even though this comparison to the most widely used and strongly competitive technology (of print) is available to almost all learners and presumably is already in the experience of the designers, the latter manage neither to think about it nor to discuss it in their papers: a quite extreme lack of reflection.

All the above common weaknesses of e-learning research are only about establishing whether anything worth noting happened at all (cf. did they get to the moon?), never mind how. For insights into process, rather than establishing the existence of a product, we need open-ended studies and measures. Here the characteristic weakness is not exactly comparability but whether the selected individuals or cases are representative, and so likely to illuminate other cases. Again that isn't a problem in classic engineering approaches: if there's only been one trip to the moon, then that is the case to study. Any one of the 100 biggest companies would be a reasonable choice for a case study of big business. But why would these three students selected for an in depth qualitative study be illuminative of other students? The experimental world too selects a few to stand for all, and their answer to this question is random selection. Random however does not mean "convenient for the researcher". A problem with many studies is that qualitative work is a much bigger, not lesser, burden on the participant than simply giving five minutes to fill in a questionnaire or 50 minutes to do an experiment, so it is even harder to get a random sample. The consequence is, that such studies are likely to focus only on students who are quite exceptionally compliant to the researcher's (or teacher's) wishes. Such studies seldom give an account of the participant selection process, still less discuss why we should think that those participants give any insight into most learners.

Of course this matters much less if open-ended methods are combined with a second pass of comparable measures, but without the second pass they simply raise issues but settle none, in fact they do not even give any reason to regard them as likely to be important. Without the second phase, we never know if the issue is general across the situation, or limited to the one or two cases/learners looked at. E.g. if two students in a class of 120 say they'd prefer to work straight through a double lecture without a break do you act on this "feedback" or survey the whole class on this specific question?

Conclusion

Above all, I think my approach is pragmatic: I'm ready to try anything that has genuine promise or better, has shown itself to produce interesting results in e-learning research, and I apply this test to both traditions I was trained in like experimental physics, and those I was not. What I'm most against is just applying a method because that's what the researcher always does, or was trained in: researcher-centered studies rather than learner-centered ones.

References

Bell Labs "Arno Penzias -- Nobel Prize" http://www.bell-labs.com/user/apenzias/nobel.html (visited 9 March 2006)

Brown, M.I., Doughty,G.F., Draper, S.W., Henderson, F.P. and McAteer, E. (1996) "Measuring Learning Resource Use." Computers and Education vol.27, pp. 103-113.

Carroll,J.M. & Rosson,M.B. (1995) "Managing evaluation goals for training" Communications of the ACM vol.38 no.6 pp.40-48

Denzin & Lincoln (1994) Handbook for qualitative research (London: Sage)

Denzin & Lincoln (1998) The landscape of qualitative research: Theories and Issues, volume 1 (Sage: London)

Draper,S.W. (1996) "Observing, measuring, or evaluating courseware: a conceptual introduction" ch.11 pp.58-65 in LTDI implementation guidelines (ed.) G.Stoner (LTDI, Heriot Watt University: Edinburgh)

Draper,S.W. & Brown, M.I. (2004) "Increasing interactivity in lectures using an electronic voting system" Journal of Computer Assisted Learning vol.20 no.2, pp.81-94

Draper,S.W., Brown, M.I., Henderson,F.P. & McAteer,E. (1996) "Integrative evaluation: an emerging role for classroom studies of CAL" Computers and Education vol.26 no.1-3, pp.17-32

Hake, R.R. (1991) "My Conversion To The Arons-Advocated Method Of Science Education" Teaching Education vol.3 no.2 pp.109-111 http://www.physics.indiana.edu/~hake/MyConversion.pdf

Hake,R.R. (1998) "Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses" Am.J.Physics vol.66 no.1 pp.64-74 http://www.physics.indiana.edu/~sdi/ajpv3i.pdf

Kearsley,G. (1994) The theory into practice database http://tip.psychology.org/ (visited 14 March 2006)

Kuhn,T.S. (1962/1970) The structure of scientific revolutions (The University of Chicago press: Chicago)

Parlett, M.R. & Hamilton,D. (1972/77/87) "Evaluation as illumination: a new approach to the study of innovatory programmes" in
[a] (1977) D.Hamilton, D.Jenkins, C.King, B.MacDonald & M.Parlett (eds.) Beyond the numbers game: a reader in educational evaluation (Basingstoke: Macmillan) ch.1.1 pp.6-22.
[b] (1987) R.Murphy & H.Torrance (eds.) Evaluating education: issues and methods (Milton Keynes: Open University Press) ch.1.4 pp.57-73

Penzias,A. (2005) "Arno Penzias -- Autobiography" http://nobelprize.org/physics/laureates/1978/penzias-autobio.html (visited 9 March 2006)

Phillips, R. (in press) Peer Mentoring in UK Higher Education

Rosenthal,R. & Jacobson,L. (1968, 1992) Pygmalion in the classroom: Teacher expectation and pupils' intellectual development (Irvington publishers: New York)

Scriven,M. (1967) "The methodology of evaluation" pp.39-83 in Tyler,R.W., Gagné,R.M. & Scriven,M. (eds.) Perspectives of curriculum evaluation (Rand McNally: Chicago)

Tinto,V. (1975) "Dropout from Higher Education: A Theoretical Synthesis of Recent Research" Review of Educational Research vol.45, pp.89-125.

Tinto,V. (1982) "Limits of theory and practice in student attrition" J. of Higher Education **53** no.6 pp.687-700