"... go back over it till you got a place where you could believe your own facts ... and then go on again from that particular place till you had the whole thing properly believed and not have bits of it half-believed or a doubt in your head hurting you like when you lose the stud of your shirt in bed."

Flann O'Brien *The third policeman*

Improving deep learning with MCQs and EVS

(MCQ = Multiple Choice Questions
EVS = Electronic Voting Systems)

Steve Draper, Glasgow University

www.psy.gla.ac.uk/~steve/talks/evs8.html

SALT@ Swansea 23 Nov 2011

Today's menu

A. Introduction
B. Several linked questions for each concept
C. The effect of Confidence ratings
D. Questions that ask about reasons, not facts
E. Learner-authored questions
F. Catalytic assessment: what is it?
G. Catalytic assessment: Evidence of learning gains
H. Catalytic assessment: How/why does it work?
I. Conclusions

Part A:

Introduction

Intro

One tactic for teaching is to use EVS. I'll show later some evidence of big learning gains from EVS (e.g. reducing first year dropout from 20% to 3%).

However EVS almost always revolves around MCQs

The traditional UK view of both students and staff of is that MCQs = pub quizzes = shallow rote learning of trivial facts.

This talk is organised as an evidence-based rebuttal of that belief; so it is about how MCQs can be the heart of teaching that promotes deep learning, and often these methods use EVS.

Intro 2

Contempt for MCQs is misplaced IF they have been professionally designed. (L1 psych)

The real reason is that it is easier to get away with a poorly designed exam question if it is an essay title or calculation-based "problem".

However professionally designed MCQs e.g. in US textbooks, while they are discriminating tests of knowledge, do mostly only test factual knowledge.

This is true of Geertje's Elimination Testing and Gardner-Medwin's CBM too.

Educationally more important is promoting "deep" learning, whatever that means.

Deep and shallow learning

It's obvious from the metaphor that deep=good, and shallow=bad. Or is it?

Some of the work covered here seems to imply that students can get top marks, yet don't understand apparently basic things. Does this mean (in non-MCQ exams):

- Exams are only testing shallow learning, whether or not they use MCQs?
- Or that there are several kinds of thing we want our students to know, but we only examine them on one kind?

Warmup questions (1)

Show of hands / EVS: decide which of the options you are going to vote for when I ask:

Is Swansea west of Glasgow? Yes / No

Warmup questions (2)

Show of hands / EVS: decide which of the options you are going to vote for when I ask:

Approximately how many HEIs are there in the UK (according to the HE Statistics Agency)?

33, 66, 99, 132, 165, 198, 231, 264

Warmup questions (3)

Show of hands / EVS: decide which of the options you are going to vote for when I ask:

Which shows deeper learning for a student in a year 1 physics course?: Getting correct:

a)qualitative reasoning about a circuit of 2 bulbs, a battery and a switch?

b) Solving Kirchoff's law equations for the circuit?

Warmup questions (4)

Show of hands / EVS: decide which of the options you are going to vote for when I ask:

In a class, seating positions in the lecture hall are randomly assigned to students, who must then use that seat for the whole course. A later analysis of overall course grades showed:

1. No effect of seat position
2. A statistically significant effect of seat position of nearly a grade from best to worse position.

Warmup questions (5)

Show of hands / EVS: decide which of the options you are going to vote for when I ask:

Which of these 2 methods as the better effect in creating bonding and "integration"?

a)A cheese and wine party for students and staff to meet socially (not attendance taken)

b)A 2 day reading party which for students is compulsory and they have to pay for, involving physical exercises, and talk presentations.

Warmup questions (6)

Show of hands / EVS: decide which of the options you are going to vote for when I ask:

Which of these 2 courses, matched for discipline, type of students, size, etc. led to LESS dropout?

a) Had a welcome tour of the city
b) Started work on day 1, no introduction, straight into project groups with a tutor, each group to present to the rest of the course on Friday

Nominal goals, and underlying goals

Simple-minded people want something and just do it.
Real life however also has many things that are pursued indirectly.

•Taking the dog for a walk
•Going to a club, going dancing, ...
•Golf for business contacts, or exercise
•Hill climbing (why not just walk up and downstairs; or take a helicopter to the top)
•Dinner parties (for food? or conversation?)

Later in the talk, you'll see that setting a question that only asks for a simple, shallow answer can nevertheless actually be the best way of getting deep learning to occur.

Part B:

Several linked questions per concept

A learner can get an MCQ right when they don't really understand the concept it is trying to test

Linked questions per concept

A student can get a 4-option MCQ right by chance 25% of the time.
And even with no guessing, getting one example of a concept right often does not predict that that student can apply it in a different question (transfer).

Mark Russell: superior tactic of linked questions for diagnosis.
"Using an electronic voting system to enhance learning and teaching" Mark Russell (2008) Engineering Education vol.3 no.4 pp.58-65

N ≈ 71 students
Three questions testing same underlying concept of pressure gradient (how pressure varies with depth in water).

Mark Russell (1)

Qu. 1 The pressure gradient (dp/dz) in a fluid ...

a) is always zero
b) is always negative
c) is always positive
d) can be either depending on the fluid
e) is equal to the gauge pressure

See figs.3-9 in:

Mark Russell (2008) "Using an electronic voting system to enhance learning and teaching" Engineering Education vol.3 no.4 pp.58-65

Mark Russell diagrams

The Mark Russell case as numbers

	Q1	Q3	Q2	1&2	1&3	2&3		All 3
% correct	83%	63%	58%	50%	51%	?		36%

The argument is that the conjunction of questions is a more powerful test of understanding than any one by itself.

Part C:

Confidence ratings

Asking about confidence

Hunt (1982) (in an artificial experiment) showed that participants who first chose an answer and then had to indicate a confidence level (*sure / not sure*) learned about 20% faster than those who just chose an answer.

(This general issue is sometimes called "metacognition": when the learner isn't just a recorder of information but reflects on their learning and may modify their learning activity because of this.)

Gardner-Medwin's CBM (Confidence Based Marking; or Certainty Based Marking) is a direct application of this.

Confidence-based marking

Basically, the learner gets an MCQ, and must first select one of the answer options, and then rate their confidence. The mark they earn is a product of the two: of their confidence multiplied by the positive or negative score for their answer.
So being uncertain doesn't score high, being very sure but wrong is heavily penalised.

Certainty level:	C=1	C=2	C=3	No reply
Mark if correct:	1	2	3	0
Penalty if wrong:	0	-2	-6	0

Students need practice to get a feel for the system, but then results are good. Extensive website, and published papers on this. http://www.ucl.ac.uk/lapt/index.htm

Part D:

You want understanding? Just test for reasons

Assertion-reason questions

Asking not about the truth of facts, but the reasons for a true fact.
- The question states a fact
- The response options list reasons
- All of these may be true
- All should have been in the course being tested
- => Then recognition will not help the student.

CAAC (Computer Assisted Assessment Centre) website advice on MCQ design:
http://www.caacentre.ac.uk/resources/objective_tests/index.shtml

Questions about reasons (sky)

The night sky is dark because:

A. The Universe is infinite in size

B. The Universe is expanding

C. The Universe is made up of, mostly, Dark Matter

D. The Universe has existed for only a finite amount of time

E. The distribution of stars is not smooth, but fractal

The weatherman on the news predicted that thunder and lightning was on its way. Why would it be a better idea to put your hood up to keep yourself dry than to use an umbrella?

a) It might be windy so the umbrella could break.
b) The metal on the umbrella's handle and spokes could conduct electricity if lightning struck it.
c) The metal on the umbrella handle could rust in the heavy rain.
d) The umbrella could block your view from any potential hazards

You must be careful to keep a can of deodorant away from fire because it can explode. This is because:

1. the heat from the fire increases the pressure in the can by giving the atoms more energy.
2. the heat from the fire decreases the atoms' energy, also decreasing the pressure in the can.
3. If the can leaks, the chemicals react badly with the fire.
4. the aerosol's particles join together and solidify so there is no longer enough room in the can.
5. The can contains liquid, and when liquid boils the pressure and/or volume increases enormously.

MCQs as learning objectives

Given that the exam is to be about learners giving reasons, not facts, you can set the work using MCQs as learning objectives. The learners must then research what the right answers are, and why.

Jaye Richards did this in a school class: a "Jigsaw" design, where the learners, not the teacher, did the exposition.

Jaye Richards' L-design

(trialled with 12 year olds in a general science course)

- Start each block with a set of MCQs designed to:
  - ❑Get them interested in what they are about to learn
  - ❑To act as learning goals: what they must find out
- Don't tell them the right answers
- Apportion the MCQs (and response options) among the groups
- Each group, over next few periods, researches & creates a presentation (learner authored answers / SGC) on:
  - ❑what the right answer is,
  - ❑why each response option is right or wrong
  - ❑Physical demonstrations
- Presentations by each group to whole class (Jigsaw)
- Retest on the original MCQs

What's powerful in Jaye's design?

- As in learner authored questions, it effectively gets them to give reasons, not rote memory
- Getting students to teach each other, and not the teacher doing exposition.
  (Jigsaw design. Betty Collis' Student generated content)
- Using (EVS) questions to define the learning agenda for these student-teachers (EBL)
- Using (EVS) questions to motivate the learning agenda Cf. "pre-lab, pre-lect" methods; constructivism, …

Part E:

Learner-authored questions

Learner authored questions

This is another powerful teaching tactic.

Basic idea:
Students have to design a test MCQ (best done by small groups) complete with reasons why each response option is right or wrong (which can then be used for automatic feedback for those attempting the question).

Have to aim for questions that discriminate (splits class).

Why is this effective? The factual question requires them to generate reasons ….

Learner authored questions (2)

This has been done numerous times, if you search the literature.

- In multiple disciplines e.g. accountancy, ...
- For credit, or not.
- With a promise to use the best MCQs written by students in the final exam, or not.

For more references see Draper (2009), and Nicol (2007)

PeerWise

There is a free software service that allows you to do this on any scale, including huge classes. (Currently, its web page says about 200 HEIs (worldwide) have an account with them.)

Students create MCQs and add them to a bank, where their whole class can use them. There are mechanisms for student reviews of how good each question is.

http://peerwise.cs.auckland.ac.nz/
The central website

http://www.psy.gla.ac.uk/~steve/localed/innovs.html#pw
Short introduction to it by me

PeerWise (2)

It will only work well if you make the authoring a compulsory assignment (unless you have a class of 1,000; in which case the 5% of volunteers will produce a critical mass of questions the whole class benefits from).

Could (like Andy Sharp) make them author MCQs as part of EVS-enhanced presentations they give to the rest of the class.

They will need a bit on instruction on what makes a good question i.e. not too hard or too easy.

PeerWise has systems for students to rate other students' questions, as well as to answer them.

Part F:

Catalytic assessment: what it is

"Catalytic assessment" is my catch-phrase for questions that may look like tests, but whose important mathemagenic (learning-generating) effect is hidden in the learner.

Catalytic assessment

"Catalytic assessment" is a catch phrase for questions that may look like tests, but whose important mathemagenic (learning generating) effect is hidden in the learner.
They are simple factual questions on the surface, but they make the learners think; and the result is far deeper understanding.

For the full argument see my paper on this:
"Catalytic assessment: understanding how MCQs and EVS can foster deep learning" British Journal of Educational Technology vol.40 no.2 pp.285-293

[This ref. is on the handout sheet]

Linking to the learner's own questions

Behind the notion of "catalytic assessment", and its focus on how learners may notice a bug in their understanding, and then correct it, is:—
The general point that learners bring with them, not a blank mind (though you can bore it into blankness) but their own ideas, questions, worries, prior conceptions.
Failing to address these means your 'teaching' will be unconvincing to the good students, irrelevant and incomprehensible to the bad.

That lies behind the success for elementary physics teaching of "brain teasers'" play with everyday experiences.

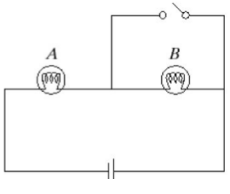
Mazur's Brain teasers

The point is to provoke debate, internal and between peers.
Cf. Socratic questioning, and "catalytic assessment"

Remember the old logo or advert for Levi's jeans that showed a pair of jeans being pulled apart by two teams of mules pulling in opposite directions. If one of the mule teams was sent away, and their leg of the jeans tied to a big tree instead, would the force (tension) in the jeans be:

- half
- the same
- or twice what it was with two mule teams?

From Mazur



When the switch is closed, the brightness of bulb A:
1. Increases.
2. Remains unchanged.
3. Decreases.
4. Cannot be determined from the information given.

Peer Instruction: Mazur Sequence

1. Concept question posed (brain teaser)
2. *Individual Thinking*: students given time to think individually (1-2 minutes)
3. Students provide individual responses
4. Students receive peer feedback – poll of responses presented as histogram display
5. *Peer Discussion*: students instructed to convince their neighbours that they have the right answer.
6. Retesting of same concept
7. Students provide individual responses (revised answer)
8. Students receive feedback – poll of responses presented as histogram display
9. Lecturer summarises and explains 'correct' response

Part G:

Catalytic assessment: Evidence

Some evidence about how MCQs / EVS can be educationally successful.

Hake

Hake (1998) published a survey of 62 courses (6,542 students) all studying the same subject, all using the same standardised test, and using it both pre- and post-.

He graphed the mean gain on each course against whether or not it had used the method of "Interactive engagement".

See fig. 1 in:

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 66(1), 64-74

Hake's results

Mazur

Crouch & Mazur (2001) published an analysis of 10 years of Mazur's MIT course.

Again, the standardised pre- and post-test.

He concludes he has doubled the amount of learning, but the graph suggests that really, he tripled it.

See fig.2 in:

Crouch, C.H. and Mazur, E. (2001), "Peer Instruction: Ten years of experience and results" American Journal of Physics 69, 970-977

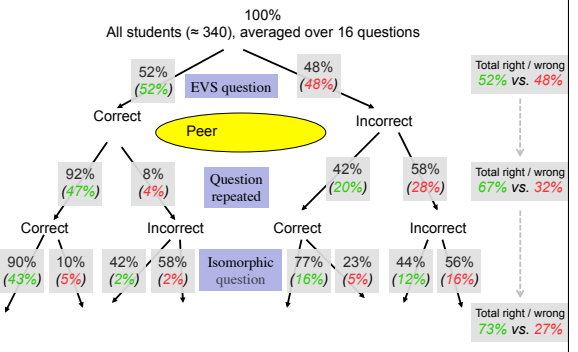
Mazur's gains

Smith et al. 2009 paper in Science

- IE / Mazur type method, but in level 1 Genetics, not physics
- Re-test was not only the identical question, but also another similar (isomorphic) one.
- Even when no-one knew the right answer, many students learned from the peer discussion (for 15 of 16 topics)
- Biggest improvement on the more difficult questions
- Delayed benefit in the sense that some got the isomorphic one right even if persisting in the wrong answer for the repeated question.

Went from 52% correct to 72.52% correct averaged over 16 qus.
(7.4% got worse; 28% better)

Redrawn from: **Smith et al. 2009 paper in Science**



Not the philosopher's stone, ...

So the Mazur recipe doesn't always work, in the sense that every learner improves their understanding after every discussion on every question.
Some learners seem to get worse.
Some questions seem to make a majority get worse.

But the overall pattern is strongly: a noisy random walk tending towards greater understanding in every class.

Here's some data from the first year a lecturer tried it, in a new subject with questions she wrote herself. (Her following year's data were better.)

Summary stats over the 58 questions on one course, each repeated after peer discussion	
88%	Percent of questions with some net gain
15%	Average net gain (pre to post, as % of class)
32%	Average normalised gain (pre to post as % of those who got it wrong and could be improved) *[2/3rds of students fail to improve]*
93%	Percent of qus. with right answer winning in the end
16%	Percent of qus. where right answer comes from behind to 1st place
57%	Percent of qus. where more than one option attracts and gains votes after discussion
57%	Average (mean) proportion of changed votes that are in the right (vs. wrong) direction *[Lots of churn]*

Part H:

Catalytic assessment: How/why does it work?

Miyake and "constructive interaction"

We can understand Hake's and Mazur's demonstrated practical educational successes in terms of the theory developed in developmental psychology of how peer interaction promotes individual's conceptual advances.

Miyake (1986) got researchers round her lab to discuss their understanding of sewing machines.

Detailed analysis of the conversations showed that this was NOT teaching, yet both did advance their conceptions.

Christine Howe's work (1)

Long series of studies on peer interaction causing conceptual development.

Good selected paper:
Howe, C.J., Tolmie, A, and Rogers,C. (1992)

To get the effect, you need to work on the setup:

Peers with different prior beliefs
Elicit commitment to their personal view in advance e.g. write their view, then show peers this opinion.

Christine Howe's work (2)

- Benefit is delayed (e.g. 4 weeks)
- Final conceptions are different in solo than group interviews
- More advanced child ALSO advances still further I.e. it is NOT information transmission
- "not agreement but private conflict resolution"

⇒ Mechanism is metacognition
(Howe, McWilliam, Cross 2005)

Michelene Chi: Self-explanations

Chi, in a career-long series of studies, showed that in a tutorial what causes the most learning is NOT necessarily when the tutor explains things i.e. gives the explanation, the reasons; but when the learner is induced to generate, to produce reasons. (Learners it seems have to do it for themselves, not have it done for them by their nanny / tutor.)

This is what peer interaction, as in Mazur's method, does; as do some other methods e.g. authoring MCQs (and their explanations). Chi's work gives strong experimental evidence that this is as good as or better than most personal tutoring by an expert: or to put it another way, the best expert tutors elicit explanations, they do not give them.

(Of course, we all enjoy being the expert, like me here showing off to you. But it isn't the best thing for your understanding.)

So:

So more effective teaching is achievable.

And it can be achieved with MCQs, asking questions of various types.

It isn't really the surface format of the question (e.g. Bloom category / level), but what **kind of thinking** it elicits in the learner.

Good catalytic questions are harder to write than they look. But friends have done it; and there are some question banks online.

Put simply

It's about getting learners to
a) Consider whether they really understand this stuff (confidence self-assessment);
b) Generate reasons to help decide ("self-explanations")

Most people just don't think, so they don't learn very well: certainly, they don't work on their understanding, only on performing adequately. Get them to think, and the learning improves in quality and quantity.

Put another way, a learner needs to …

"... go back over it till you got a place where you could believe your own facts ... and then go on again from that particular place till you had the whole thing properly believed and not have bits of it half-believed or a doubt in your head hurting you like when you lose the stud of your shirt in bed."

Flann O'Brien *The third policeman*

Part I:

Conclusions

Hestenes

- He created the FCI test for understanding (≈ like Mazur qus.)
- The Hake, and Crouch&Mazur results are gains in FCI scores (pre/post).
- Exam results don't go up
- But the FCI has huge face validity with the teachers: they just cannot bear that their students can't do these items.
- What does this mean?
- It is understanding; it is deep learning. It is the kind of learning that is wanted by teachers but not normally tested for
- Yet (Mazur's students) demonstrated in exams complex equation-solving which is a core disciplinary skill.
- So: there seem to be some other core skills, ...

- Deep learning (I suggest) is about having multiple, not one, skill.
- Shallow learning can be complex, yet inadequate.

Deep and shallow learning

It's obvious from the metaphor that deep=good, and shallow=bad. Or is it?

Some of the work covered here seems to imply that students can get top marks, yet don't understand apparently basic things. Does this mean:

- Exams are only testing shallow learning, whether or not they use MCQs?
- Or there are several kinds of thing we want our students to know, but we only examine them on one kind?

A place to stop

Discussion QUESTIONS

1.What learning aims are you testing for?

2.What attributes do you assume your students should have, but you don't test for?

3.Can you see how to use MCQs to promote the kind of learning that matters to you?

For the slides, handout etc. see:

http://www.psy.gla.ac.uk/~steve/talks/evs8.html