Tony Gardner-Medwin

Nancy Curtin

UCL

Imperial College London

The Skill of Knowing how Reliable your Knowledge is

CONFIDENCE-BASED MARKING (CBM)
.... in formative & summative assessment

• **What it is**
• **How it relates to knowledge and learning**
• **What you can expect if you use it**
    .... Experience from UCL & Imperial
• **How you can set about it**
• **Issues of fairness , good practice, reliability, validity**
• **Evaluation & dissemination - discussion**

www.ucl.ac.uk/lapt

Confidence-Based Marking (CBM)

Knowledge is not just about getting it right
  - you must know when you are getting it right
  - and you must indicate if you aren't sure
These are life skills – relevant in every discipline

LEARNING-ORIENTED ASSESSMENT must develop self-monitoring of the quality of work (Sadler, Knight)

Confidence-Based Marking (CBM)

Knowledge is not just about getting it right
  - you must know when you are getting it right
  - and you must indicate if you aren't sure
These are life skills – relevant in every discipline

LEARNING-ORIENTED ASSESSMENT must develop self-monitoring of the quality of work (Sadler, Knight)

• *The simplest, most fundamental self-monitoring relates to right/wrong answers*
• *A student who can discriminate between reliable and unreliable answers deserves more credit than one who cannot, even if they get the same number of answers correct.*

Words that might describe a student answer

Argued	Believed	Careless
Checked	Confident	Correct
Delusion	Hesitant	Ignorant
Justified	Knowledge	Lucky
Misconception	Misinformation	Reflected upon
Rigorous	Supported	Uncertain
Understood	Unlucky	Wrong

Which deserve reward ?

Words that might describe a student answer

Argued	Believed	Careless
Checked	Confident	Correct
Delusion	Hesitant	Ignorant
Justified	Knowledge	Lucky
Misconception	Misinformation	Reflected upon
Rigorous	Supported	Uncertain
Understood	Unlucky	Wrong

How can we reward these and encourage better learning?

Words that might describe a student answer

Argued	**Believed**	Careless
Checked	**Confident**	**Correct**
Delusion	**Hesitant**	**Ignorant**
Justified	**Knowledge**	Lucky
Misconception	Misinformation	**Reflected upon**
Rigorous	**Supported**	**Uncertain**
Understood	Unlucky	Wrong

CBM rewards confidence in <u>correct</u> answers and reservation about <u>wrong</u> answers – placing a premium on reflection, justification & understanding.

What is knowledge ?

✓ knowledge
✓ uncertainty
○ ignorance
✗ misconception
✗ delusion

decreasing confidence in what is true, increasing confidence in what is false

Knowledge is intimately related to confidence, or degree of certainty

To philosophers: knowledge is <u>justified</u> <u>true</u> <u>belief</u>

We need to treat it as such:
- **to reward not just truth**
- **but justification**
- **and discrimination of levels of certainty or belief**

What is CBM ? - *experience it at www.ucl.ac.uk/LAPT*

The UCL (LAPT) Confidence-Based Marking scheme

… applied to individual answers that will be marked right/wrong
…   e.g. T/F, MCQ, EMQs, Numerical, Simple text

Degree of Certainty	1	2	3
Score if Correct	1	2	3
Score if Incorrect	0	-2	-6

Best marks obtained if use when :

Probability correct is	**< 67%**	**67-80%**	**>80%**
Odds are	**< 2:1**	**> 2:1**	**> 4:1**

With CBM you must think about justification

You gain:

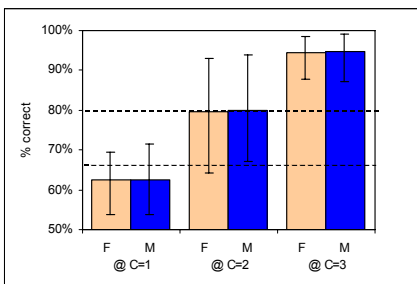<u>EITHER</u> if you find justifications for high confidence

<u>OR</u> if you see justifications for reservation.
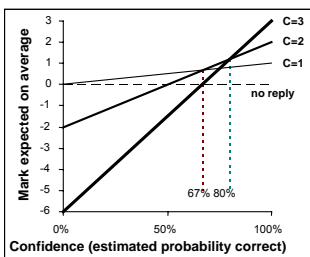
The best C level is the one with the highest graph :

C=3 when P(correct) >80%

C=1 when P(correct) <67%

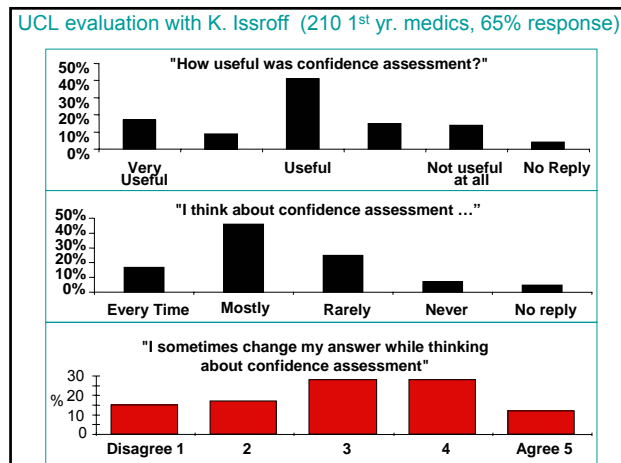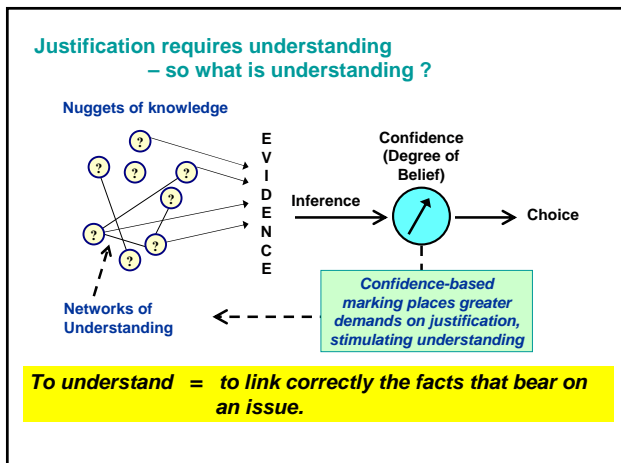### How well do the students discriminate reliability?

Bars show range including 90% of students

[331 students, 500 T/F exam questions, 2001]

Students adopt confidence-based marking very easily

<u>Principles that students seem readily to understand</u> :-

- If you can't judge the reliability of your knowledge (either over- or under-estimating it) – it will cause problems for later learning
- confident errors are far worse than acknowledged ignorance and are a wake-up call (-6!) to pay attention to explanations
- expressing uncertainty when you are uncertain is a **good thing** . Students familiar with fixed negative marking (±1) particularly like the option to acknowledge low confidence and avoid risk of penalty.
- thinking about the basis and reliability of answers can help tie bits of knowledge together (to form "networks of understanding")
- checking an answer and rereading the question are worthwhile
- judging reliability of ideas should be a routine study habit

**Justification requires understanding
– so what is understanding ?**

Nuggets of knowledge

E
V
I
D
E
N
C
E

Confidence
(Degree of
Belief)

Inference → Choice

Networks of
Understanding

*Confidence-based
marking places greater
demands on justification,
stimulating understanding*

*To understand = to link correctly the facts that bear on
an issue.*

UCL evaluation with K. Issroff (210 1st yr. medics, 65% response)

"How useful was confidence assessment?"

Very Useful | Useful | Not useful at all | No Reply

"I think about confidence assessment …"

Every Time | Mostly | Rarely | Never | No reply

"I sometimes change my answer while thinking
about confidence assessment"

Disagree 1 | 2 | 3 | 4 | Agree 5

- Trial exercise
- NAC Experience at ICL

Imperial College
London

**The skill of knowing how reliable your
knowledge is**

Nancy Curtin

6 Sept 2005, Improving Student Learning Symposium

Imperial College
London

Confidence based marking

Background:

Medical students:

~350 intake

1997 to date

Imperial Med Sch = CXWMS + Marys merge

New curriculum "Tomorrow's Doctors"

mainly Yr1 & 2 of 6 yr course

Imperial College
London

CBM in LAPT: Self-assessment

Availability of LAPT:

Desk-top of College computer

Down load to home via disk or web

Alerting students:

In hard-copy Course Guides

Links in Students' Intranet

Yr1 "How to study" session

Individual Lecturers

bush-telegraph

Imperial College London   CBM in LAPT: **Self-assessment**

Sources of material:

Exercises from individual lecturers *(few)*

Specimen Examination papers

Student summer projects financed by Med Sch
Learning Resources Budget *(popular)*

"Generic" pre-1997 question banks

UCL exercises

Imperial College London   CBM in LAPT: **Self-assessment**

Evaluation:

Staff-Student Liason Committee

Not in SOLE

Use statistics

Questions seen

100,000

50,000

0

1998-9   2000-1   2002-3   2004-5

Academic Year

Imperial College London   CBM in **Formative** assessment:

1st trial in Jan 05 for Yr1 students

also 1st on-line Formative

Results:

Excellent student participation: >90%

staff enthusiastic

Disappointing performance.

Lack of practice?

Not "exam conditions"?

Imperial College London   CBM in **Formative** assessment:

Future: Jan 06

Yr1 Formative to be repeated

New Yr2 Formative scheduled

Interest in extending use to other Yr1 & Yr2 Formatives

Interest among Clinical Examiners

Importance of Practice with CBM:

UCL students had had several compulsory online tests.

ICL students were encouraged to practice online, but many had not done much.

Formative tests:
at UCL (150 Qs, paper), Imperial (45 Qs, online)

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%

CBM scaled: chance=0%

0%   20%   40%   60%   80%   100%

No. correct: scaled so chance=0%

adj, lin'd

UCL students – compare online self-assessment with exams

100%

90%

80%

70%

60%

50%

% correct

@C=1   @C=2   @C=3

F M   F M   F M   F M   F M   F M
(i-c)   (ex)   (i-c)   (ex)   (i-c)   (ex)

Mean +/- 95% confidence limits, 331 students
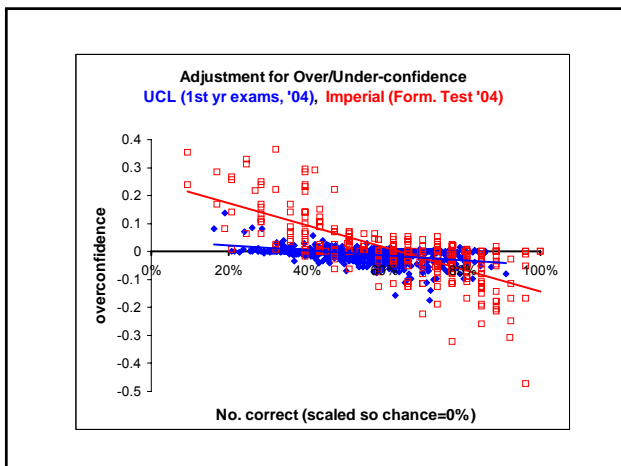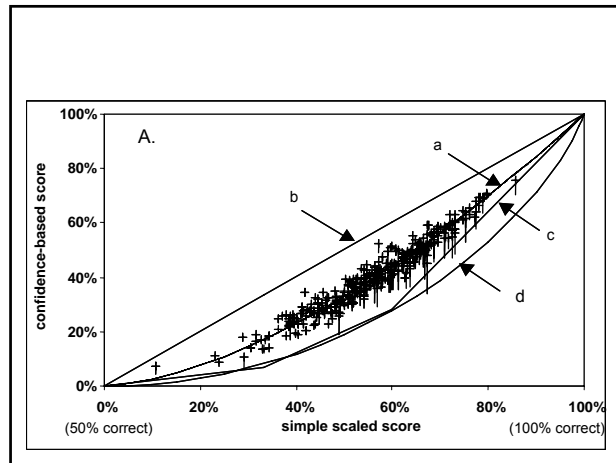
Slide 1

Are there students with poor calibration in exams?

- and how should one handle them ?

Significantly overconfident in exam: **2 students (1%)**

e.g. **50%** correct @C=1, **59%** @C=2, **73%** @C=3

Significantly underconfident in exam: **41 students (14%)**

e.g. **83%** correct @C=1, **89%** @C=2, **99%** @C=3

[UCL Yr 1 2000/01 (500TF Qs)]

Maybe one shouldn't penalise such students

Adjusted confidence-based score:

Mark the set of answers at each C level as if they were entered at the C level that gives the highest score.

mean benefit = 1.5% ± 2.1% (median 0.6%)

Slide 2

Slide 3

Adjustment for Over/Under-confidence
UCL (1st yr exams, '04), Imperial (Form. Test '04)

Slide 4

Issues about using CBM in exams?

- We introduced CBM in part because students said they would prefer it, following their experience using it for revision & practice
- **Transparency**, simplicity & motivation in the CBM scheme are important
- **Formative practice** is important, and possibly the most valuable thing
- Experience with formative practice doesn't require **exam board** approval!
- **Reliability & Validity** of exam data have been increased
- **Passmarks** in exams can be set initially using % correct, as previously
- You can use **Virtual Learning Environments** & **Optical Mark Readers**

Slide 5

Reliability and Validity of Confidence-based exam marks

Exam marks are determined by:

1. the student's knowledge and skills in the subject area
2. the level of difficulty of the questions
3. chance factors - how questions relate to details of the student's knowledge and how uncertainties resolve (luck)

(1) = "signal" (its measurement is the object of the exam)

(3) = "noise" (random factors obscuring the "signal")

Confidence-based marks improve the "signal-to-noise ratio"

A simple & convincing test of this is to compare marks on one set of questions with marks for the same student on a different set (e.g. odd & even Q nos.). High correlation means the data are measuring something about the student, not just "noise".

Slide 6

Marks scaled: 0%=chance 100%=max

The correlation, across students, between scores on one set of questions and another is higher for CBM than for simple scores.

But perhaps they are just measuring ability to handle confidence ?

No. CBM scores are better than simple scores at predicting even the <u>simple</u> scores (ignoring confidence) on a different set of questions. This can only be because CBM is statistically a more efficient measure of knowledge.

Cronbach Alpha (standard psychometric measure of 'reliability')

On six exams (mean ± SEM, n=6):

α = 0.925 ± 0.007 using CBM

α = 0.873 ± 0.012 using number of items correct

- The improvement (P<0.001, paired t-test) corresponds to a reduction of the random element in the variance of exam scores from 14.6% of the student variance to 8.1%.
- This almost halves the length of an exam for equivalent reliability

---

What tools are available to help?

- **The website** has tools & 1000s of example Qs :
-                 www.ucl.ac.uk/lapt
- **Publications:** evaluation, statistics, pedagogy, concerns, review
- **Authoring tool:** full range of Q formats (TF, MCQ, EMQ, Number, text)
- **Manual** available for full options – conditional, random features etc.
- **Conversion** tools (e.g. from WebCT quiz format)
- **Syntax checker** (currently requires installation)
- **Full character sets**, HTML options available
- Use software from the UCL site, or install it locally. E.g. if you put your own exercise file at www.my_url.js , try www.ucl.ac.uk/lapt?www.my_url
- No new Q types are required – easy to adapt existing Qs or WORD files
- At UCL we can help with editing, VLE, OMR, any technical problems.

Confidence-Based Marking

    = **Certainty based Marking**

        = **True-Knowledge based Marking**

           = **Marking to stimulate understanding**

What evaluation Qs would you like us to ask?

Examples :
Rate on a scale from 1=strongly disagree to 5=strongly agree
- CBM encourages me to think more carefully about answers.
- If I am confident and wrong, then I deserve a penalty (negative mark).
- CBM rewards gamblers rather than those who understand the topic well.

CONCLUSIONS

- **CBM is core "Learning-Oriented Assessment"**
  - **to enhance self-monitoring**
  - **usable with existing/new exercises**
  - **encourages deeper thinking**
  - **rewards & trains an important generic skill**
- **Is about knowledge – not attitude or personality**
- **Improves the reliability of exams**
  - **(with formative practice of course)**
- **At UCL we will help you implement it**

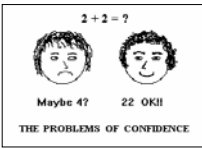
We fail if we mark a lucky guess as if it were knowledge.

We fail if we mark delusion as no worse than ignorance.

www.ucl.ac.uk/LAPT

Personality, gender, culture issues: real or imagined ?

Does confidence-based marking favour certain personality types?

2 + 2 = ?

Maybe 4?   22 OK!!

THE PROBLEMS OF CONFIDENCE

- *Both underconfidence and overconfidence are bad!*
- *'Correct' judgement is clear, desirable and trainable*
- *No gender or ethnic differences are evident (at least after practice)*
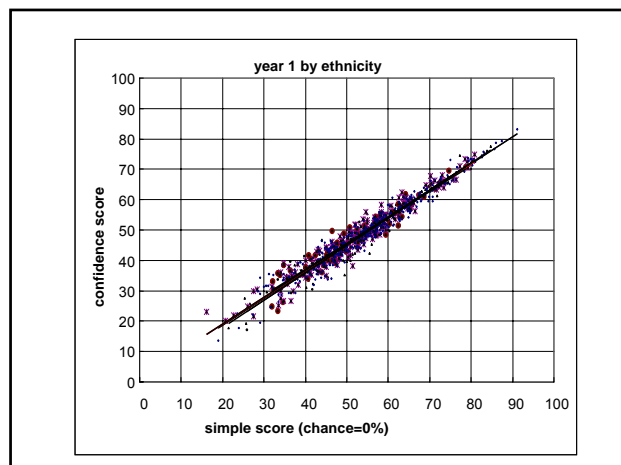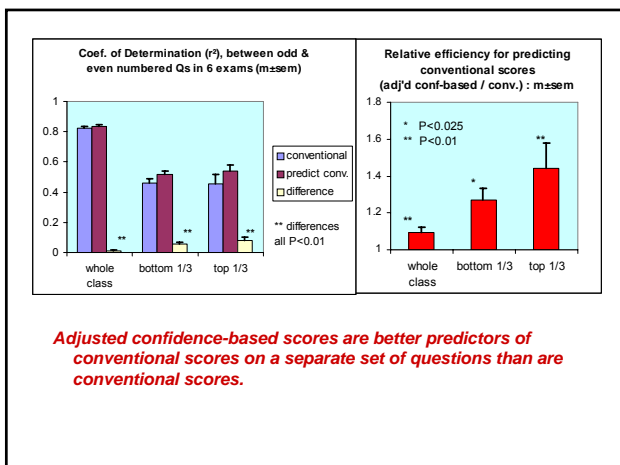- *Students with confidence problems: this is the way to deal with it!*
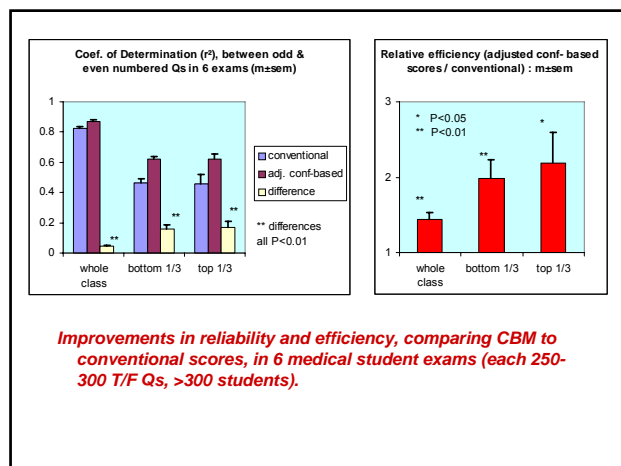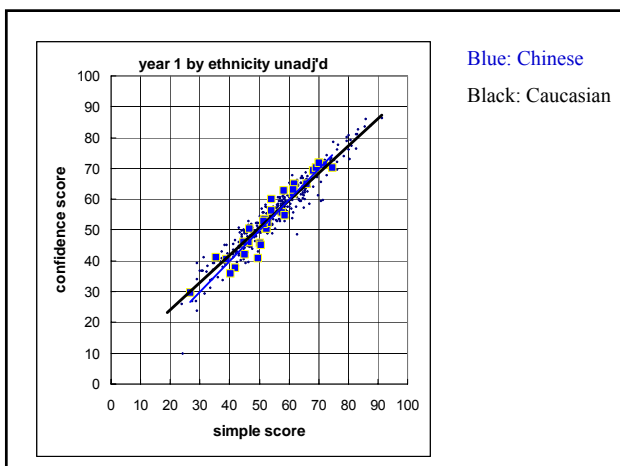

Confidence-Based Marking (CBM)

A student who can discriminate between reliable and unreliable answers deserves more credit than one who cannot, even when each gets the same number of answers correct.
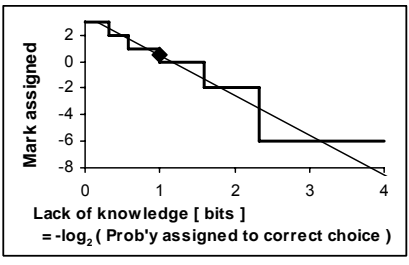
CBM marks each answer according to the student's degree of certainty that the answer is correct.

| Degree of Certainty :	C=1 (low)	C=2 (mid)	C=3 (high)	No Reply
Mark if correct:	1	2	3	0
Penalty if wrong:	0	- 2	- 6	0

CBM rewards students who can reflect to the point that they can either :
(a) justify confidence in answers, or (b) identify reasons for reservation.
*It is <u>motivating</u> – always rewarding honest reporting of confidence.*

Blue: Chinese

Black: Caucasian

Improvements in reliability and efficiency, comparing CBM to conventional scores, in 6 medical student exams (each 250-300 T/F Qs, >300 students).

Adjusted confidence-based scores are better predictors of conventional scores on a separate set of questions than are conventional scores.

Marks as a function of lack of knowledge defined by information theory, for T/F answers

Lack of knowledge [bits]
= -log$_2$ (Prob'y assigned to correct choice)

Rational and Irrational marking schemes
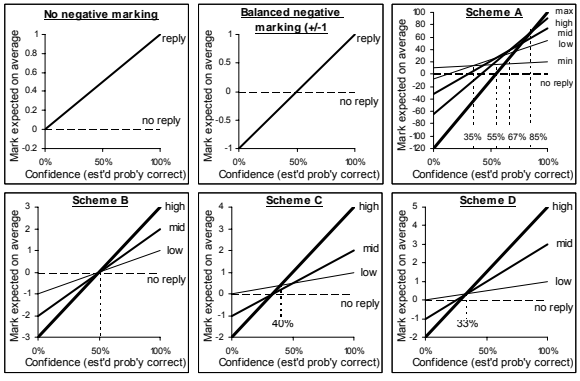
Breakdown of credit and variance due to uncertainty
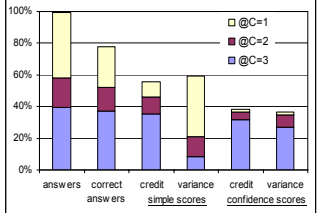
Simple scores *(scaled conventional scores) were scaled so chance gives 0%, total knowledge 100% (equivalent to +1 for correct, -1 for incorrect, 0 for omission).*
- 65% of the variance came from answers at C=1, but only 18% of the credit.

Confidence scores: these give less weight to uncertain answers; uncertainty variance is then more in proportion to credit, and was reduced by 46% (relative to the variation of student marks)