Comparing Class Scores in GCSE Modular Science

JASON WELCH, County High School, Leftwich, Cheshire, UK

Abstract

Multiple choice tests are used widely in education and elsewhere. The results of these tests contain information both about the students' knowledge and their ability to guess the answers. This paper describes the use of Bayesian statistical techniques to attempt to 'remove' the guess-work from the results in order to obtain information about the students' underlying knowledge based on our prior knowledge about the structure of the test. The resulting mathematical model allows fair comparisons of the levels of knowledge of groups of students in schools and highlights the flaws in the common practice of analysing these scores using simple averages. It also allows more specific comparisons to be made that are not possible using averages. These comparisons can then inform teaching practice.

Introduction

A multiple choice test provides the student with a number of options from which they are to select the correct answer *e.g.*

The Milky Way is a …

A galaxy

B solar system

C universe

D star

Using such a test to assess knowledge can be problematic not least because the person being tested could guess the correct answer without any understanding of the topic. The literature on multiple-choice testing is wide-ranging but can be broadly categorised into four areas: question writing, administration of tests (electronically), scoring systems and results analysis. The work comes predominantly from higher-education (especially in medicine, law, economics and IT) with contributions from statistics and psychology.

The use of multiple-choice over other forms of assessment clearly depends strongly on question effectiveness and this has been studied extensively, often using Bloom's taxonomy (Bloom (1956)) *e.g.* Simkin and Kuechler (2005).

Electronic test administration allows both new response types such as confidence assessment (Gardner-Medwin (1999)), and the opportunity for adaptive questioning in which questions are selected dynamically based on performance (see Kurhila *et al.* (2001)).

Scoring and analysis often attempts to address the problem of guessing. A good overview of scoring methods is the often-cited paper by Bush (1999), with several others such as Angoff and Schrader (1981), Muijtens *et al.* (1999), and Gardner-Medwin (1999), comparing or analysing specific methods in more detail. The use of negative marking techniques to discourage guessing has attracted a significant literature in psychology (see for example Bar-Hillel *et al.* (2004)) with students often reporting that such methods are 'unfair'. Burton (2001) analyses a similar situation to the one considered here to develop a measure of test reliability, including the effects of question selection.

This paper considers only single-response number-right scoring, the situation found in UK GCSE modular science examinations, although the analysis applies to any similarly-structured multiple-choice test. The choice of delivery method (paper or screen) is not relevant to the analysis and the problem of writing good questions is not considered. This paper is focused specifically on the problem of inferring individual or class knowledge levels (to the extent that such a quantity can be said to exist) from their scores on the test.

Raw Data

In UK GCSE modular science examinations, marks are awarded for correct responses only; there is no 'negative marking' or other system of penalising guesswork. The raw scores so obtained are then adjusted to give a universal modified score (UMS). This mapping is in general non-linear and depends on the performance of students across the country. Empirically the mapping is linear in the middle and lower scores, but penalises the highest scoring students as shown in Figure 1. These adjusted scores obscure the true performance so when we are interested in student knowledge and not headline A*-C percentages, 'raw' scores (where available from the examination board) should always be used in comparing performance and are used throughout this paper. We would like to process the data to find

out to what extent the students' scores were influenced by lucky guessing and how much the scores reflect their knowledge of the subject.

<Figure 1 here>

Bayesian Methods

Bayesian methods are statistical techniques that can help in problems such as this, where the data (test scores) are confused by some kind of external influence (guessing). These methods work by incorporating other information we have about the problem (being taught by the same teacher, for example). A Bayesian method incorporates this prior information in a fair way, allowing the data to modify and eventually overwhelm the prior knowledge as the quality and quantity of data increases. A wide literature on these methods exists and interested readers are directed to Box and Tiao (1992) for an introduction.

As an example we will consider the problem of analysing a single student's score on a test.

Basic Model (I_0)

We will study the results of a student taking a test in which there are L questions with R responses.

Each student is assumed to *know* the answers to k questions, and to correctly guess g questions. The student will therefore score $s = k + g$. This is a simplified model because students will often make informed guesses by eliminating some of the 'distractors'. We are unable to use this information however, because it depends on the individual student and is not identified by the testing method.

We would like to infer the student's knowledge, k, from their score, s. A Bayesian analysis allows us to obtain not a single "right" answer for k, but a probability distribution for k indicating our state of knowledge about what k might be. The Bayesian approach is to interpret probabilities such as $pr(x)$ as a 'state of knowledge' about x, and to use Bayes' Theorem to manipulate these quantities. Quantities that are part of the model but are not directly of interest are 'integrated out', effectively adding the contributions of all possible values. In our case, we will integrate g out of the problem, so that our solution takes into account all possible combinations of guesses.

The mathematical details are included for readers familiar with Bayesian methods but other readers may wish to skip directly to the results below. The analysis yields:

$$pr(k|s)=\sum_{g}\frac{pr(s|g,k)\,pr(g|k)\,pr(k)}{pr(s)}$$

The next step in a Bayesian analysis is to assign *prior* distributions, that is, decide our state of knowledge about quantities before receiving the data. Without the scores, an analyst who does not know the class would have no preconceptions about *k* and assume each possible value to be equally likely. We assign a uniform prior distribution in [0,*L*] on each *k*. *pr(g|k)* is a binomial distribution, selecting *g* correct from *s–k* guesses with probability *1/R*. *pr(s|gk)* is zero unless *g=s–k*, mathematically *pr*(*s*|*g*,*k*) = (*s-(g+k)*), selecting one term from the sum. Ignoring the constant factors *pr*(*k*) and *pr*(*s*) and taking the logarithm yields:

$$\log pr(k|s)=\log(L-k)!+(L-s)\log(R-1)-\log(s-k)!-\log(L-s)!-(L-k)!$$

Results using I_0

Figure 2 shows the results obtained using this formula for R=4, L=24. Denoting the maximum of the distribution as $k=\hat{k}$ we see that

- $k\hat{\rightarrow}0$ for small *s* consistent with pure guess-work;
- $k\hat{\rightarrow}s$ for large *s* where the student is very knowledgeable having little reliance on guess-work; and
- $k\rightarrow\hat{s}-\hat{g}$ with $2\le\hat{g}\le4$ elsewhere, indicating an expected contribution of two to four marks from guess-work for average students.

These results are in line with what we would have expected from ‘common sense’ and we can now develop the analysis to more complex examples.

<Figure 2a> <Figure 2b>

<Figure 2c> <Figure 2d>

Class Group Model (I_1)

We now consider a *class* of *N* students. We assume they have been set for ability and have been taught by one teacher. This means that we expect, before they take the test, that there will be some consistency in the scores. We may not believe this very strongly, but when the data arrive our Bayesian method will allow the data to overwhelm these prior beliefs if there is enough evidence for doing so. Mathematically this is done by allowing correlations between the students' *k*'s.

<Figure 3 here>

To model such a correlation, our prior knowledge of the value of *k* for any student in the class is taken to be centred on a middle value *m* and having a certain width. (We have used an exponential decay model in which *pr(k=m+1) / pr(k=m)* = constant, *f*, as shown in Figure 3.) Figure 3 represents our expectation of the spread of students' *knowledge* before we see their scores. We expect they will cluster around a central value with a few students knowing more or fewer answers than this.

We are now interested in finding a value of *m* which gives an indication of the level of knowledge of the class as a whole. We use a Bayesian analysis that takes into account their scores, the likely extent of guesswork as described in the previous section, and the expected correlation between their levels of knowledge. This analysis produces not a single answer for *m*, but a probability distribution for *m*, allowing us to state our level of belief in any particular value of *m* given all the information to hand.

The mathematical analysis is set up as before:

$$pr(m|s) = \int_f \sum_g \sum_k \frac{pr(s|g,k,m,f)\,pr(g|k,m,f)\,pr(k|m,f)\,pr(m|f)\,pr(f)}{pr(s)}\,df$$

where bold type is used to denote vector quantities such as $\mathbf{s} = (s_1, s_2, \ldots, s_N)$, the N students' scores. We 'integrate out' the model parameters that are not of direct interest as before, namely *f*, **g**, and **k**, thus taking account of all possible values in a fair way. In the case of **g** and **k**, as these are discrete variables, the integration is a sum.

We now need to encode our prior knowledge in the probability distributions in the equation above. Standard Bayesian procedures exist for assigning priors on certain quantities (see Box and Tiao (1992) for more details). These suggest a uniform prior in [0,L] on m (independent of f) and a prior on f proportional to f^{-1} in some suitable range $0<f<1$. In order to perform the calculations on computer we will discretise f to some suitable resolution in the range. $pr(\mathbf{g}|\mathbf{k},m,f)$ is the binomial distribution discussed before. $pr(\mathbf{s}|\mathbf{g},\mathbf{k},m,f)$ again selects from the sums those g for which $s = g + k$. $pr(\mathbf{k}|m,f)$ is proportional to $f^{-|k-m|}$ for each k independently, modelling our exponential prior on k centred on m with fall-off rate, f.

Putting all of these ideas together and ignoring the constant $pr(\mathbf{s})$ for the moment yields:

$$pr(m|s) \propto \sum_f \frac{Z^{-N}(f,m)}{(L+1)} \frac{Z_f}{f} \prod_i \sum_{k_i} f^{-|k_i-m|} \frac{(L-k_i)!(R-1)^{L-s_i}}{(s_i-k_i)!(L-s_i)!R^{L-k_i}}$$

where i is an index running across the students in the group.

Although the quantity of interest is the probability distribution for m, we would like also to obtain $pr(\mathbf{s})$ as this is the probability of obtaining the data based solely on the use of this model. We will call this quantity the 'evidence' for the model and we will use it later to compare models. Mathematically we require all other distributions to be normalised so that $pr(\mathbf{s})$ is simply the normalising constant for the final distribution. We have denoted this normalisation for $pr(\mathbf{k}|m,f)$ as $Z(f,m)$, and for $pr(f)$ as Z_f in the chosen discrete range.

Test Group Results using I_1

Three 'toy' datasets were generated based on N = 20, L = 24, R = 4. In 'toyA', a group expected to achieve 'A' grades, each student scored 21. In 'toyG', expecting 'G' grades, students scored between 6 and 9. In 'toyC', expecting 'C' grades, students scored between 9 and 14.

The results are shown as Figure 4; each plot shows $pr(m|\mathbf{s})$, the resulting or *posterior* distribution for m, given the scores. The results show the patterns expected; the model predicts 'toyA' knew 20 answers, 'toyC' knew 9 answers, and 'toyG' knew 3 answers. This is sensible given the basic model where perhaps 2-4 marks are scored from guessing. The correlations have had the effect that the

spread of scores in the group is explained by a range of knowledge around *m*, and by guess-work. Once the guess-work has been accounted for, only a small range of knowledge is required to explain the results, hence the posterior is narrower than the original range in the data.

<Figure 4 here>

Real Group Results using I_1

Three real datasets for a GCSE science module test on 'Inheritance' were then analysed based on L = 24, R = 4. Data from three classes who took this test were analysed; class 3, N=27; class 6, N=21; and class 8, N=19. The classes are ability sets in numerical order in Year 10, predicted mostly 'C', 'D' and 'F' grades respectively. The results are shown as Figure 5.

In these results we see a similar pattern to the 'toy' datasets. In class 8 however, the posterior is flatter and wider. This is due to a larger spread of scores in this class. The posterior shows the model's lack of ability to distinguish uniquely between spread of ability and spread of luck in guessing. The advantage of the Bayesian approach is that the lack of uniqueness is *quantified* (has an associated probability level) and can be used in fair comparisons. We shall see later that, with additional information, we *can* in fact go some way to distinguishing between spreads in ability and in guesswork.

<Figure 5 here>

Comparison of Models

The class group model I_1 incorporates our assumption that the scores will be correlated (grouped around *m*) based on the fact that the class has been set for ability and taught by the same teacher. The Bayesian method allows us to compare this assumption with the assumption I_0 that each student's performance was individual (that the correlations are not real).

The 'evidence' for a model is *pr*(model|data). Given two models we can compare them using:

$$\frac{pr(I_1|s)}{pr(I_0|s)} = \frac{pr(s|I_1)\,pr(I_1)\div pr(s)}{pr(s|I_0)\,pr(I_0)\div pr(s)}$$

$$= \frac{pr(s|I_1)}{pr(s|I_0)}$$

where I_0 and I_1 are the models (between which we assume no preference in advance), and ***s*** are our data. The normalising constant for *pr(k|s)* in the basic model (single student) and *pr(m|**s**)* in the class group model is associated with *pr*(**s**|$I_{0/1}$) above, the evidence for the model mentioned previously. The ratio measures to what extent the data tell us we should prefer one model over the other. Typically these ratios are large powers of 10 because *pr*(**s**) is a product of perhaps 30 independent data, each with probability of perhaps 0.3, resulting in evidence ~$0.3^{30}=10^{-16}$.

Applying this analysis to the 'toy' and class datasets reveals overwhelming evidence in support of the correlations proposed in the whole-class model (see Table 1). In other words, the data provide strong evidence for the supposed correlations between students' knowledge. This is weakest in class 8 where the spread of scores suggests more individuality in the students' performances.

<Table 1 here>

Using Additional Prior Information

Some examination boards' papers have still more structure which we can use as prior knowledge in our model. We will consider the case where the questions in the paper are grouped into sections *of increasing difficulty*. This means that we expect scores to *decrease* as we progress through the sections of the paper for all but the highest ability student. We will take the example of three sections of eight questions each, four answer choices, with each section increasing in difficulty.

To begin, consider the example of two students A and B whose scores in the three sections were (3,4,5) and (7,4,1) respectively. Although equally graded with a raw score of 12, most analysts would say that student B *knew* more correct answers than student A. With a low score in the easy section, student A is unlikely to have known five answers in the most difficult section; she probably guessed

more successfully than did student B. We might suggest that student A's result is reasonably consistent with knowing little of the subject, while student B's result shows sound basic knowledge (but perhaps not advanced knowledge).

In order to model this mathematically we will separate **s**, **k** and **g** into three parts s_1, s_2, s_3 etc. one label for each section of the paper, section one being the easiest. The analysis follows exactly as above until faced with assigning a prior to $k_{1,2,3}$. We factorise the prior as follows:

$$pr(k_1, k_2, k_3) = pr(k_3|k_2, k_1)\, pr(k_2|k_1)\, pr(k_1)$$

We assume all scores in section one to be equally likely *a priori* setting the final term to be constant in, in our case, [0,8]. We assume further that the correlation between k_1 and k_3 is made through k_2, therefore $pr(k_3)$ is independent of k_1 directly and we take the prior to be:

$$pr(k_1, k_2, k_3) = pr(k_3|k_2)\, pr(k_2|k_1)/9$$

The posterior distribution is now three-dimensional, a 9 x 9 x 9 grid. Each set of *k*'s will have its own probability assigned. While this may be difficult to visualise, we can simplify matters later by adding the contributions of *k*-combinations that make the same total score *i.e.* we can look at $pr(\sum k_i|s)$ instead of $pr(k_1,k_2,k_3|s)$ – mathematically, projecting the posterior onto lines of constant $k=\sum k_i$.

To proceed we must select a mathematical function to model the correlation between the k_i.

Choice of Prior for k_2 and k_3

Since the sections are of increasing difficulty we expect *a priori* that the *k*'s will not increase from k_1 through k_3. There should be quite a large penalty against $k_3 > k_2 > k_1$ so that only overwhelming data would allow us to accept that a student actually knew many more answers in section three than in section one. For $k_1 > k_2 > k_3$ however we have no *a priori* preference for additional structure as a good student will have $k_1 = k_2 = k_3 = 8$, while a weaker student may have $k_1 = 8 >> k_2 > k_3 = 0$ (a sound grasp of the basics but weak application and synthesis skills in Bloom's (1956) terminology).

The behaviour we want is shown in Figure 6. A large value of k_1 allows probability to 'leak' into larger k_2's. Small k_1's pull most of the prior for k_2 into $k_2=0$. The rate at which the probability decays to higher k_2 and the extent of leakage for middle k_1's can be controlled by a parameter. The same function can be applied to the prior for k_3 given k_2.

A possible mathematical form for this is:

$$pr(k_2|k_1) = Ae^{\frac{k_2(k_1-8)}{\quad}}$$

where A is a normalising constant and determines the extent to which probability leaks to high k_2. When $k_1 = 8$ the prior is uniform in k_2 as required; when $k_1 = 0$ there is an exponential decay to higher k_2 controlled by .

<Figure 6a> <Figure 6b>

<Figure 6c> <Figure 6d>

We now face the typically Bayesian dilemma of how to set , called the 'regularisation constant'. Small represents a strong belief that students weak in one section will know nothing in the harder sections, therefore favouring *k*-combinations such as (6,0,0) over (5,1,0). Large says that students' knowledge is less hierarchical. It would be possible to put a prior on and then integrate over all values in some reasonable range, raising the question of what is 'reasonable'. On the other hand, different analysts may legitimately *choose* different reflecting their views on students' knowledge. A discussion of the merits and technical details of both approaches can be found in MacKay (1999) and Bretthorst (1988). We will take a pragmatic, empirical approach for the following reasons.

It is possible to have different regularisation for k_3 and k_2. One argument for this could be that, since there are more middling students than there are exceptional students, we should have a higher decay rate in k_3. This would be further justified if we could say that section three questions were *much* more difficult than section two, *i.e.* that the increase in difficulty was non-linear. Another argument says that many students will *guess* better in section two than in section three by narrowing down choices based on partial knowledge *i.e.* the distractors are easier to identify in section two. This would require

a similarly sharp cut-off into section two. Experience suggests that the data will be good enough to decide this, and we propose to choose values for the two that maximise the evidence for the model.

Class Group Section Score Model (I_2)

We must now build a model in which both the correlations between students and between sections are considered. In the notation used previously we have:

$$pr(m, \quad |s) = \int_f \sum_g \sum_k \frac{pr(s|g,k,m,f,}{}$$

where plain text letters now denote vector quantities such as $m \equiv m_j = (m_1, m_2, \ldots, m_J)$ and bold denotes matrix quantities such as $S \equiv s_{ij} = (s_{11}, s_{21}, \ldots, s_{N1}, s_{21}, s_{22}, \ldots, s_{NJ})$ for N students answering L questions in each of J sections on the paper with R responses from which to choose. We thereby recover a posterior distribution for m, which we can maximise with respect to as described, and use to assess the level of ability of the class. In our example above, J=3 sections, L=8 questions and R=4 choices.

Prior information comes in the form of correlations between students' scores and between section scores for each student individually. The former correlation is encoded as before using the exponential decay prior for each k_{ij} around that section's m_j, with fall-off rate f_j. This prior is applied to each section independently allowing different f's for each. This models the students' tendency to score similarly, due to the common setting and teacher. The latter correlation is encoded as above, although it is applied to the m_j rather than to the individual student's set of k's. This way we link the group's scores in each section according to their increasing difficulty, expecting $m_j < m_{j-1}$.

Using our previous formula and inserting the additional correlations we can now write:

$$pr(m, \quad |s) \propto \prod_j pr(m_j | m_j - 1,$$

Substituting our form for $pr(m_2|m_1)$ and using a scale parameter prior for in a suitable range, as we did for *f*, yields **Equation 1**:

$$pr(m, \quad |s) \propto \frac{1}{9} \prod_{j=1}^{J} A_j e^{\frac{m_j (m_{j-1} - J)}{\quad}}$$

A computer programme was built to implement this formula and run for the example case J=3 sections, L=8 questions, R=4 choices. A copy of the programme is freely available from the author.

Test Group Results using I_2

The programme was run against simple test data to check expected outcomes (see Table 2). Tests 1-3 show agreement with a common sense view. In test 4, the data provide some evidence for knowledge in section 3 *greater* than in section 2, against the prior, and the posterior begins to reflect this.

<Table 2 here>

The 'toy' datasets used earlier were adapted to provide section scores in keeping with the group's expected grade. 'ToyA' consisted in 20 students scoring 21, adapted so that each scored 8,7,6. 'ToyG' consisted in 20 students scoring between 6 and 9; combinations used suggested extensive guess-work *e.g.* 2,0,6. 'ToyC' consisted in 20 students scoring between 9 and 14; mostly sensible combinations such as 7,3,2 were used. The results, shown in Figure 7, can be compared with Figure 4 from the earlier model. We have a sharper posterior distribution for 'toyA' where all students scored the same, but a flattened and right-skewed result for both 'toyC' and 'toyG'. The additional information provided by the section scores has lent some credibility to a higher level of group knowledge (and reduced guessing) because students generally scored higher in earlier sections.

The results are, however, somewhat sensitive to changes in and, in particular, when is very small, the posterior can change by several percent. We chose to select based on the evidence, but it turns out that reducing *always* increases the evidence. To see why this is, consider for simplicity both 's to be equal. They then come out from under the sum in Eqn. 1 as a constant multiplier.

Making smaller by a large factor then increases the value of each cell of the posterior by that factor, but should be offset by a worse fit to the data (smaller likelihood value). However, the model has another escape route available – the value of *f* can increase to broaden *pr(k)* and continue to fit the data. This was checked by obtaining the joint posterior *pr(m, , f | s)* and noting the maximum in *f* for different . In each case, as decreased, the best *f* increased accordingly to fit the data; consequently the value in each posterior cell increased and therefore so did the evidence. This tells us that our model is over-complicated with more parameters than are needed to explain the data.

<Figure 7 here>

An alternative and slightly less strict prior (which comes from the empirical form for the charge density in an atomic nucleus) was tried, with the algebraic form:

$$pr(m_2|m_1) = \frac{A}{1+e^{(m_2-m_1)}}$$

This time the density is broadly constant with increasing m_2 until $m_2 = m_1$ when an edge is reached and the probability rolls-off to zero (Figure 8). The precise shape of this roll-off is controlled by . By setting $= 1 / (m_1 + a)$ we get a roll-off that is faster when m_1 is small, as required. This also removes the nuisance scale parameter, , in favour of a location parameter *a*. Unfortunately this prior either gives lower evidence than the original prior (with smooth roll-off), or gives results that permit unsatisfactory *m*-combinations such as (6,6,6) (with higher-value , sharper cut-offs). Analysts who feel this combination is not unsatisfactory may however choose to use this form of prior.

To remove the parameter altogether, a simpler functional form was tried. This uses a linear, rather than exponential, decay rate with a gradient dependent on the previous section score. A gradient of $(m_{i-1} - 8)/32$ was used, where m_{i-1} is the previous section's *m*-value. This gives a flat prior for m_i when $m_{i-1}=8$ and a reducing, hard cut-off value for reducing m_{i-1}. Although this is initially disturbing as it fully forbids some *m*-values in some cases (*i.e.* sets some *pr(m)=0*), in practice the results showed little difference from those of the exponential prior with reasonable . In projection, the results were indistinguishable to within 1%. This prior was then used as the basis for all further analysis.

<Figure 8 here>

Real Group Results using I_2

Figure 9 shows the results of applying the final model to the real class data analysed earlier.

<Figure 9 here>

Comparing Figures 9 and 5 we see that the posterior distributions for all classes have been skewed to the right. We conclude that the individual section scores show more evidence of knowledge than the first model predicted; they could show this, for example, by containing more scores like 8,4,0 than 4,8,0. Class 6 has the weakest skew – a look at the scores shows several guesswork combinations, such as 5,3,5 and 2,1,3. Class 8 is interesting in that the data include students with scores of 8,6,6 and 8,5,5 as well as others with 5,0,1 and 6,0,2. Teachers would probably recognise a genuinely wide range of knowledge in this group and this is reflected in the flatter posterior for this group.

The results from using the linear prior on *m* consistently produce significant increases in evidence over the exponential models in many cases. In some cases the evidence is smaller but only by factors of O (1). This check confirms that the choice of prior is satisfactory.

Comparing the new model (I_2) to the original model (I_1, with only total score, not section scores) is not simple because we have essentially different data. This means that our evidence procedure is not valid. However, a little manipulation shows that:

$$\frac{pr(I_1|s)}{pr(I_2|s)} = \frac{\sum_{s_0} pr(s|s_0, I_1)\, pr(s_0|I_1)}{pr(s|I_2)}$$

where *s* represents the new data (in sections), and s_0 represents the original data. In the sum, any dataset s_0 in which the section scores do not add to the total score has first term zero. Only the original dataset meets this criterion. We therefore retain the evidence ratio, but we modify the original model by a 'weight factor' equivalent to the probability of getting the section scores given the full scores, $pr(s|s_0)$. Model I_1 does not know about sections, so all valid section scores are equally probable. If there are *W* ways of making up a given total score from section scores, then the weight factor on that score is $W / (L+1)^3$ (*L* being 8 in the present case).

Table 3 shows the evidence values for various data and model combinations. As discussed earlier, model I_1 is overwhelmingly supported over the uncorrelated model, I_0 (see row B/A). Similarly, model I_2 is supported over *its* uncorrelated model (say I_0', as shown in row F/E). Model I_2 is also superior to model I_1, despite the additional complexity, by a factor of at least 10 for these data.

<Table 3 here>

Table 4 shows a comparison between a traditional analysis of the raw scores and the results found above. The traditional approach of looking at means of total scores would congratulate the students or teacher of class 8 and vilify that of class 6 and class 3 (being only a little ahead of class 8). Even an enlightened school considering variance of the raw scores will not be able to draw valid, clear conclusions as the variance is large as a result of guessed answers. There is only weak evidence in the traditional analysis for any difference between the groups. The Bayesian analysis has plenty to say: class 3's level of knowledge matches expectations of 'C' grade students; class 8 has done well, although there is a wider range of knowledge within the group; class 6 has under-performed. In addition, the full posterior can be used to ask more detailed questions such as "what is the probability that class 8 only knew answers in section one?" (0.18), or "what is the probability that class 6 knew fewer than two questions over sections two and three?" (0.55). Answers such as these can be used to inform teaching practice (to focus revision effort after a 'mock' examination, for example).

<Table 4 here>

Conclusion

Current analysis of GCSE modular science scores is often unscientific. The use of UMS scores under-represents the contribution of high-achievers. The use of simple averages (even with consideration of variance) does not give a true picture of a group's performance because no allowance is made for guess-work. The Bayesian method offers a probability distribution for true performance providing the opportunity to ask more complex questions of the data. This information can be used both to inform individuals' teaching, and to assist Heads of Science in identifying effective teaching methods for different groups. With the expected continuation of this form of assessment in the new 2006 science specifications, this technique could be a useful tool for use in schools in the future.

References

Angoff, W. H. & Schrader, W. B. (1981) A Study of Alternative Methods for Equating Rights Scores to Formula Scores, *Research Report*, Educational Testing Service.

Bar-Hillel, M., Budescu, D. & Attali, Y. (2004) Scoring and Keying Multiple Choice Tests: A Case Study in Irrationality, *Discussion Paper Series dp370*, Center for Rationality and Interactive Decision Theory, Hebrew University, Jerusalem.

Bloom, B. S. (ed.) (1956) *Taxonomy of Educational Objectives, Handbook 1: Cognitive Domain* (New York, McKay).

Box, G. E. P. & Tiao, G. C. (1992) *Bayesian Inference in Statistical Analysis* (London, Wiley Classics Library).

Bretthorst, G. L. (1988) *Bayesian Spectrum Analysis and Parameter Estimation* (London, Springer-Verlag).

Bush, M. (1999) Alternative Marking Schemes for Online Multiple Choice Tests, in proceedings of *Seventh Annual Conference on the Teaching of Computing*.

Kurhila *et al.* (2001). Bayesian Modelling in an Adaptive Online Questionnaire for Education and Educational Research, in proceedings of *PEG2001*.

MacKay, D. J. C. (1999) Comparison of Approximate Methods for Handling Hyperparameters, *Journal of Neural Computation*, **11**, 1035-1068.

Muijtjens, A.M.M., van Mameren, H., Hoogenboom, R.J.I., Evers, J.L.H. & ven der Vleuten, C.P.M. (1999) The effect of a ‘don't know’ option on test scores: number-right and formula scoring compared, *Medical Education*, **33**, 267-275.

Simkin, M. G., Kuechler, W. L. (2005) Multiple-Choice Tests and Student Understanding: What Is the Connection? *Decision Sciences – The Journal of Innovative Education*, **3**(1), 73-98.