# EMPIRICAL PREDICTION OF THE MEASUREMENT SCALE AND BASE LEVEL 'GUESS FACTOR' FOR ADVANCED COMPUTER-BASED ASSESSMENTS

Don Mackenzie & David O'Hare

# Empirical Prediction of the Measurement Scale and Base Level 'Guess Factor' for Advanced Computer-based Assessments

Don Mackenzie & David O'Hare

Centre for Interactive Assessment Development
University of Derby
Kedleston Road
Derby
DE22 1GB

D.Mackenzie@derby.ac.uk

## Abstract

In our experience, insufficient consideration is often given to the way in which the questions in computer-based assessments are scored. The advent of more complex question-styles such as those delivered by the TRIADSystem (Mackenzie, 1999) has made it much more difficult to predict the distribution of possible scores and the base level guess factor than it has been for tests containing simple multiple-choice questions. For example the TRIADS drag and drop template allows each object to be allocated a different score (positive or negative) for each position as well as allowing dummy objects and dummy positions to be defined. The number of score possibilities for a random answer increases dramatically as the number of objects and positions is increased and although a 0 to 100 scoring scale is available, scores are likely to be concentrated about 'nodes' on this measurement scale. The positions of these 'nodes' will vary with the structure of the question and negative or penalty scoring may serve to 'smear' the mark distribution between 'nodes'. Many tutors may find it difficult to predict the guess factor and will not appreciate the effect that the structure of the question may have on the range and distribution of final scores achieved.

In order to demonstrate to test designers the effects of question structure and score allocation on the 'guess factor' and mark distributions, we are developing an empirical Marking Simulator. This program allows test designers/tutors to select a question type, enter the proposed structure and scores for each question then view the mark distribution and measurement scale that would result from a set of entirely random answers.

Use of the marking simulator should result in a more realistic setting of pass levels and generally enhance the quality of computer-based assessments.

## Introduction - The Marking Simulator

In three years of production of computer-based assessments on a university-wide basis, we have found that, in many cases, tutors unfamiliar with the use of objective items place insufficient emphasis on appropriate scoring methodologies for questions (Boyle et.al, 2002). There is also a general lack of appreciation of the extent to which scores achieved by guesswork alone may influence the final mark distribution from a computer-based assessment and score corrections may be made on an ad-hoc basis. The assumption by many that output from a computer-based assessment is directly comparable with that from a traditional exam-type assessment leads to the potential application of pass-marks that are too low and the risk that academic standards will be compromised unwittingly.

We are currently developing a Marking Simulator program in order to promote better question design and a more informed interpretation of the results of computer-based assessments. This program uses a random number generator to mimic the selections and operations that are required to answer a question of a specified design using a range of scoring strategies then tests the outcomes using a large number of iterations. The graphical output displays the average score and distribution of scores that would be achieved by a large number of candidates randomly selecting the answer with absolutely zero external influences, knowledge or understanding of the question.

It is important to emphasise that this paper merely deals with the fundamental mathematical properties of the question structure in an attempt to refine question designs. We make no attempt to address the arguments over the validity of guesswork or other parameters that may influence the candidates score over and above the base levels that we discuss here.

This paper is very much a report on progress, the detailed outcomes of the investigations will be published and the program made available at a later date.

## The Base Level Guess Factor (BLGF)

It is self-evident that many question types commonly used in computer-based assessments possess an inherent design weakness in as much as candidates may obtain a proportion of their score by guesswork. Most obvious of these is the very popular, simple multiple-choice question (MCQ) where the probability of gaining a score with zero knowledge is $(1/d)*100$, where $d$ is the total number of options.

There has been much discussion in the literature of the validity of various formulae to correct scores from such assessments (e.g. Burton & Miller, 2000; Ebel & Frisbie,1991; Rust & Golombok, 1999; Ryle, 1996). However the purpose of the Marking Simulator is not to validate or promote any particular method of correction but merely to provide useful data that allows the more detailed appreciation of the distribution of marks resulting from any particular question design and scoring regime.

In this paper we distinguish between the Zero Knowledge Average Score or Base Level Guess Factor (BLGF) and the Guess Factor (GF). The Base Level Guess Factor (BLGF) is the mathematical probability of gaining a score by means of random selection of answers with zero knowledge, as defined above. The Guess Factor (GF) is the probability of a real candidate gaining a score by guessing and incorporates a whole host of parameters that include not only the quality design of the question content but also the background experience, psychological character and gender of the candidate.

The concept of the BLGF is thus useful in separating the mathematical scoring properties of the question type from the properties imparted by a more variable content-design and candidate-centred element.

For MCQs the computation of the BLGF for each question is a relatively trivial task. However, some authors suggest that the simple MCQ question type is unsuitable for testing higher level skills (e.g. Boyle et.al, 2002, Huff & Sireci 2001) and may not appropriate to use for undergraduate students. Most assessment systems (e.g.TRIADS Mackenzie, 1999) now provide a much wider range of question types. Many of these are superior to simple MCQs in as much as they provide more detailed information on candidate performance and provide scoring mechanisms for grading candidates within individual questions as well as in the test as a whole.

The potential for fine tuning of scoring in these question styles makes the computation of the BLGF less intuitive and there are dangers that unexpected results may ensue from best intentioned but uninformed scoring regimes. This is particularly the case with assessments designed by tutors new to computer-based assessment. Accordingly, the Marking Simulator will allow tutors to view the likely effects of varying in the scoring regime for individual questions and tune it to reduce the BLGF to a minimum whilst retaining as much information as possible about candidate performance. Ultimately, the program will predict the overall BLGF and score nodes for a whole assessment containing a range of question styles.

In this paper we start by demonstrating how some of the underlying properties of commonly-used simple Multiple Response question types may be investigated using the Marking Simulator, then move to the potentially more complex Extended Matching Item question types. Once the output has been proven for these question-types the Marking Simulator functionality will be extended to cover all question-types supported by the TRIADSystem during the next year.

## MSS results from multiple-response questions (MRQ)

For the purpose of this paper, a multiple-response question is defined as a question in which the candidate is required to select two or more correct answers from a list of options. This question type encompasses simple multiple hot-spot questions where the selection is from a range of obvious hotspots. Both the number of correct answers and the number of options may vary so that this definition includes matrix style questions that may have twenty-five or more options.

Multiple response questions can be configured to run in one of three principle modes:

- Constrained selection:
  the candidate is forced to make a prescribed number of selections, usually the same as the number of correct answers;
- Partially constrained selection:
  the candidate may make any number of selections up to the number of correct answers (usually) or above the number of correct answers and less than the maximum number of options;
- Unconstrained or Open selection:
  the candidate may make any number of selections up to the maximum number of options.

Constrained selection questions force candidates to guess the answer if they do not know it. Normally, constrained selection questions will only be employed where there is a critical academic requirement that all answers should be correct in order to achieve any score at all. Where the rubric of the question indicates the number of selections to be made, but does not force that number, then BLGF parameters for this type of question will have properties that are intermediate between constrained and partially constrained configurations.

The Marking Simulator is programmed to randomly generate the appropriate number of selections in each of the modes outlined above and output the BLGF parameters for a range of negative scores on incorrect selections. The results of trial runs demonstrate that there is a regular change in both the number and distribution of the scoring nodes with increased negative scores on options. The initial number of scoring nodes is dependent upon the number of correct answers but this decreases with increased negative scores on options, with nodes in the higher scoring range (below 100%) being lost first.

The exact nature of the pattern depends upon the number of correct answers, the total number of options and upon the way in which the question is configured. Constrained questions demonstrate the smallest number of scoring nodes whereas unconstrained or open questions demonstrate the largest number of scoring nodes. Partially constrained questions demonstrate an intermediate number of scoring nodes.

The Marking Simulator also calculates the proportion of guessing-candidates likely to score on each node. A study of this is particularly instructive. It can

be seen that some configurations of multiple-response questions are particularly poor in as much as there is a greater chance of scoring a positive score by guesswork than there is of scoring zero (Tables 1 to 3).

One unexpected output from the Marking Simulator is the demonstration of the potentially erroneous interpretations that could be placed on the investigation of the candidate performance on individual items. In some configurations, using a traditional 40% pass-mark, the percentage of candidates that would appear to have passed a question by random selection is substantially greater than the BLGF. An uniformed tutor might interpret the number passing this question as an indication that the cohort had adequately understood the topic of the question. In reality the number is merely a facet of the distribution of scoring nodes and the probability of obtaining the appropriate scores by guesswork alone on a poorly constructed question.

The perception of scores achieved will depend upon whether or not guesswork is assumed. For a 40% pass-mark, the rationale for scoring the 2/5 selection so that a candidate selects 1 correct answer but admits to not knowing the other gains 50% whereas a candidate selecting 1 correct answer + 1 incorrect answer scores 40% seems logical. Marking Simulator output tables illustrating the BLGF properties for a question of this type are included below (Tables 1 to 3).

A feature of all these configurations is the relatively high negative scoring values that need to be applied to reduce the BLGF to around 10% if carry-over negative scoring is not applied. Very high proportions of candidates would appear to have exceeded a 40% pass-mark at low levels of negative scoring in the case of constrained selection and across all levels of negative scoring in the case of partially constrained selections.

The constrained example shown in Table 1 demonstrates a readily predictable BLGF of 40% where negative scoring is not implemented and less predictable 34% for the 10% penalty version. What may be unexpected however is that the number of candidates who would achieve a 40% pass in this question by random selection of answers is around 70% and that more candidates could score 50% than could score zero. Clearly would be unwise to use this question configuration unless all scores below 100% were zeroed, but even at this level 10% of candidates could pass the question by random selection of answers.
Even if the question zero were set at the BLGF value and individual scores re-scaled between BLGF and 100, the expected average score for a test full of these questions would be 20% (Residual BLGF).

**BASE LEVEL GUESS FACTOR (BLGF) PARAMETERS**

| | |
|---|---|
| Number of correct answers: 2 | Total number of options: 5 |
| Selection forced to number of correct answers | Negative scores resolved to zero |
| Number of iterations: 5000 | Residual BLGF = (((Q%-BLGF)/(100-BLGF)*100) |

| | Negative scores on incorrect options | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Parameter** | **0** | **-10** | **-20** | **-30** | **-40** | **-50** | **-60** | **-70** | **-80** | **-90** | **-100** |
| BLGF% | 40 | 34 | 28 | 21 | 17 | 9 | 10 | 9 | 9 | 10 | 10 |
| % passing Q'n at 40% | 70 | 71 | 10 | 9 | 11 | 9 | 10 | 9 | 9 | 10 | 10 |
| % passing Q'n at 40% of (100-BLGF) | 10 | 10 | 10 | 9 | 11 | 9 | 10 | 9 | 9 | 10 | 10 |
| 40% pass-mark equiv't at 100-BLGF | 64 | 60 | 57 | 53 | 50 | 46 | 46 | 45 | 46 | 46 | 46 |
| Residual BLGF assuming QScore=0 at BLGF | 20 | 15 | 12 | 9 | 11 | 9 | 10 | 9 | 9 | 10 | 10 |

**SCORE NODE LIST**

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 50 | 40 | 30 | 20 | 10 | 100 | 100 | 100 | 100 | 100 | 100 |
| 3 | 100 | 100 | 100 | 100 | 100 | | | | | | |

**PERCENTAGE OF CANDIDATES SCORING ON EACH NODE**

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 30 | 29 | 30 | 30 | 30 | 91 | 90 | 91 | 91 | 90 | 90 |
| 2 | 60 | 61 | 59 | 61 | 59 | 9 | 10 | 9 | 9 | 10 | 10 |
| 3 | 10 | 10 | 10 | 9 | 11 | | | | | | |

**Table 2 MRQ 2/5 partially constrained selection**

**BASE LEVEL GUESS FACTOR (BLGF) PARAMETERS**

| | |
|---|---|
| Number of correct answers: 2 | Total number of options: 5 |
| Open selection up to 2 answers. | Negative scores resolved to zero |
| Number of iterations: 5000 | Residual BLGF = (((Q%-BLGF)/(100-BLGF)*100) |

| | Negative scores on incorrect options | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Parameter** | **0** | **-10** | **-20** | **-30** | **-40** | **-50** | **-60** | **-70** | **-80** | **-90** | **-100** |
| BLGF% | 25 | 23 | 22 | 21 | 20 | 18 | 17 | 17 | 17 | 17 | 18 |
| % passing Q'n at 40% | 47 | 46 | 31 | 33 | 34 | 33 | 32 | 32 | 32 | 32 | 33 |
| % passing Q'n at 40% of (100-BLGF) | 3 | 2 | 3 | 2 | 3 | 33 | 32 | 32 | 32 | 32 | 33 |
| 40% pass-mark equiv't at 100-BLGF | 55 | 54 | 53 | 52 | 52 | 51 | 50 | 50 | 50 | 50 | 51 |
| Residual BLGF assuming QScore=0 at BLGF | 17 | 16 | 15 | 14 | 15 | 15 | 14 | 14 | 14 | 14 | 15 |

**SCORE NODE LIST**

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 50 | 40 | 30 | 20 | 10 | 50 | 50 | 50 | 50 | 50 | 50 |
| 3 | 100 | 50 | 50 | 50 | 50 | 100 | 100 | 100 | 100 | 100 | 100 |
| 4 | | 100 | 100 | 100 | 100 | | | | | | |

**PERCENTAGE OF CANDIDATES SCORING ON EACH NODE**

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 53 | 54 | 53 | 53 | 53 | 67 | 68 | 68 | 68 | 68 | 67 |
| 2 | 44 | 14 | 16 | 14 | 13 | 31 | 30 | 30 | 30 | 29 | 31 |
| 3 | 3 | 30 | 29 | 30 | 31 | 3 | 2 | 2 | 2 | 3 | 3 |
| 4 | | 2 | 3 | 2 | 3 | | | | | | |

**Table 1 - MRQ 2/5 Constrained selection**

**BASE LEVEL GUESS FACTOR (BLGF) PARAMETERS**

Number of correct answers: 2 — Total number of options: 5

Open selection up to 5 answers. — Negative scores resolved to zero

Number of iterations: 5000 — Residual BLGF = (((Q%-BLGF)/(100-BLGF)*100)

| | Negative scores on incorrect options | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Parameter** | **0** | **-10** | **-20** | **-30** | **-40** | **-50** | **-60** | **-70** | **-80** | **-90** | **-100** |
| BLGF% | 44 | 34 | 27 | 21 | 17 | 13 | 13 | 12 | 11 | 11 | 10 |
| % passing Q'n at 40% | 69 | 50 | 35 | 30 | 24 | 24 | 24 | 17 | 17 | 18 | 18 |
| % passing Q'n at 40% of (100-BLGF) | 20 | 19 | 17 | 9 | 24 | 24 | 17 | 17 | 17 | 18 | 18 |
| 40% pass-mark equiv't at 100-BLGF | 66 | 61 | 56 | 52 | 50 | 48 | 48 | 47 | 47 | 47 | 46 |
| Residual BLGF assuming QScore=0 at BLGF | 25 | 19 | 17 | 14 | 12 | 12 | 11 | 10 | 10 | 9 | 10 |

**SCORE NODE LIST**

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 50 | 20 | 10 | 10 | 10 | 50 | 40 | 30 | 20 | 10 | 50 |
| 3 | 100 | 30 | 30 | 20 | 20 | 100 | 50 | 50 | 50 | 50 | 100 |
| 4 | | 40 | 40 | 40 | 50 | | 100 | 100 | 100 | 100 | |
| 5 | | 50 | 50 | 50 | 60 | | | | | | |
| 6 | | 70 | 60 | 70 | 100 | | | | | | |
| 7 | | 80 | 80 | 100 | | | | | | | |
| 8 | | 90 | 100 | | | | | | | | |
| 9 | | 100 | | | | | | | | | |

**PERCENTAGE OF CANDIDATES SCORING ON EACH NODE**

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 31 | 34 | 37 | 49 | 51 | 76 | 76 | 77 | 77 | 76 | 82 |
| 2 | 49 | 5 | 12 | 4 | 18 | 21 | 6 | 6 | 6 | 6 | 15 |
| 3 | 20 | 11 | 16 | 17 | 7 | 3 | 15 | 14 | 15 | 15 | 3 |
| 4 | | 16 | 4 | 8 | 14 | | 3 | 3 | 3 | 3 | |
| 5 | | 15 | 14 | 14 | 6 | | | | | | |
| 6 | | 4 | 8 | 6 | 3 | | | | | | |
| 7 | | 7 | 6 | 3 | | | | | | | |
| 8 | | 5 | 3 | | | | | | | | |
| 9 | | 3 | | | | | | | | | |

**Table 3 MRQ 2/5 unconstrained selection**

The output for the partially constrained version of this question in Table 2 demonstrates some 10% improvement in the BLGF values. This configuration allows the candidate to admit to not knowing one of the answers by quitting the question after one selection, thus an additional scoring node at 40% is available for a 10% penalty. Pass rates are still high on uncorrected scoring but reduce to an acceptable 2% if the question scores are zeroed below BLGF then re-scaled although the residual BLGF values will contribute around 16% to the average score.

Where no hint is given to the candidate as to the number of correct answers (unconstrained) the predicted scoring will be as shown in Table 3. Whilst the BLGF value for a 10% penalty is similar to the partially constrained example, the percentage of candidates passing after BLGF correction is the highest of all three configurations there is a noticeable increase in the number of possible scoring nodes. The large number of evenly spaced scoring nodes could be seen as advantageous in increasing the level of discrimination between candidates, however in this case, the nodes represent then inclusion of relatively high numbers of incorrect selections. Thus a candidate scoring 70% (all answers selected) will have included a greater number of errors in their answer than a candidate scoring 40% (1correct + 1 incorrect). Clearly this scoring strategy is erroneous and adopting a penalty score of -40 will overcome the problem of scoring by total selection. However with a penalty score of -40, the combination of 1 correct + 1 incorrect selection scores 10% instead of the 40% in original academic brief whereas 2 correct + 1 incorrect will score 60%. The relative merit of the two scoring schemes is thus a matter of academic judgement.
In practice, it is likely that the majority of candidates will make a number of selections close to the number of correct answers in an unconstrained question and thus the BLGF parameters will be closer to that of the partially constrained example.

Overall, there is a trade-off between discrimination and BLGF. In order to design questions that will provide maximum discrimination between good and poor candidates, the maximum number of evenly distributed scoring nodes is desirable. On the other hand, the example given above illustrates that care is needed because the extra high scoring nodes may be purely a function of the candidate making a greater number of errors and this type of distribution can result in relatively high values of BLGF. Thus some limit on the number of selections allowed may be desirable. The ideal distribution will be one with the maximum number of evenly distributed scoring nodes but with the smallest proportion of students scoring on the higher value nodes by random selection.

The examples cited above illustrate that even apparently simple question configurations can have unexpected and complex scoring outcomes.

In TRIADS assessments, matrix configurations with larger number of options are possible so that 5 from 10 and 5 from 25 combinations are often used (Table 4 and Table 5).

Increasing the incorrect/correct answer ratio clearly improves the values of BLGF so that the 5 from 25 matrix question (Table 5) displays values of BLGF less than 10% even without negative scoring. Application of a negative score to incorrect selections of -20% improves the BLGF even further whilst retaining the same number of evenly distributed scoring nodes with very low proportions of guessed answers scoring on the higher value nodes. In this configuration 90% of guessing candidates score zero and 99% score 20% or less.

**BASE LEVEL GUESS FACTOR (BLGF) PARAMETERS**

| | |
|---|---|
| Number of correct answers: 5 | Total number of options: 10 |
| Open selection up to 5 answers. | Negative scores resolved to zero |
| Number of iterations: 10000 | Residual BLGF = (((Q%-BLGF)/(100-BLGF)*100) |

| | Negative scores on incorrect options | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Parameter** | **0** | **-10** | **-20** | **-30** | **-40** | **-50** | **-60** | **-70** | **-80** | **-90** | **-100** |
| BLGF% | 22 | 15 | 11 | 9 | 8 | 8 | 7 | 7 | 7 | 7 | 7 |
| % passing Q'n at 40% | 28 | 13 | 12 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 |
| % passing Q'n at 40% of (100-BLGF) | 7 | 5 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 40% pass-mark equiv't at 100-BLGF | 53 | 49 | 46 | 46 | 45 | 45 | 44 | 44 | 44 | 44 | 44 |
| Residual BLGF assuming QScore=0 at BLGF | 8 | 8 | 7 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |

**SCORE NODE LIST**

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 20 | 10 | 20 | 10 | 20 | 10 | 20 | 10 | 20 | 20 | 20 |
| 3 | 40 | 20 | 40 | 20 | 40 | 20 | 40 | 20 | 40 | 40 | 40 |
| 4 | 60 | 30 | 60 | 30 | 60 | 30 | 60 | 40 | 60 | 60 | 60 |
| 5 | 80 | 40 | 80 | 40 | 80 | 40 | 80 | 60 | 80 | 80 | 80 |
| 6 | 100 | 50 | 100 | 50 | 100 | 60 | 100 | 80 | 100 | 100 | 100 |
| 7 | | 60 | | 60 | | 80 | | 100 | | | |
| 8 | | 70 | | 80 | | 100 | | | | | |
| 9 | | 80 | | 100 | | | | | | | |
| 10 | | 100 | | | | | | | | | |

**PERCENTAGE OF CANDIDATES SCORING ON EACH NODE**

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 26 | 38 | 61 | 62 | 71 | 70 | 73 | 73 | 74 | 74 | 75 |
| 2 | 46 | 17 | 28 | 9 | 21 | 3 | 19 | 0.4 | 18 | 18 | 17 |
| 3 | 22 | 24 | 9 | 18 | 6 | 19 | 6 | 18 | 6 | 6 | 6 |
| 4 | 6 | 8 | 2 | 3 | 2 | 0.4 | 2 | 6 | 2 | 2 | 2 |
| 5 | 0.7 | 8 | 0.4 | 6 | 0.4 | 6 | 0.3 | 2 | 0.3 | 0.3 | 0.3 |
| 6 | 0.01 | 3 | 0.01 | 0.3 | 0.02 | 2 | 0.02 | 0.3 | 0.01 | 0.01 | 0.01 |
| 7 | | 2 | | 2 | | 0.3 | | 0.01 | | | |
| 8 | | 0.5 | | 0.3 | | 0.03 | | | | | |
| 9 | | 0.2 | | 0.02 | | | | | | | |
| 10 | | 0.02 | | | | | | | | | |

**Table 4 MRQ 5/10 partially constrained selection**

**BASE LEVEL GUESS FACTOR (BLGF) PARAMETERS**

Number of correct answers: 5 — Total number of options: 25

Open selection up to 5 answers. — Negative scores resolved to zero

Number of iterations: 25000 — Residual BLGF = (((Q%-BLGF)/(100-BLGF)*100)

| | Negative scores on incorrect options | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Parameter** | **0** | **-10** | **-20** | **-30** | **-40** | **-50** | **-60** | **-70** | **-80** | **-90** | **-100** |
| BLGF% | 9 | 4 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| % passing Q'n at 40% | 6 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| % passing Q'n at 40% of (100-BLGF) | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 40% pass-mark equiv't at 100-BLGF | 45 | 42 | 41 | 41 | 41 | 41 | 41 | 41 | 41 | 41 | 41 |
| Residual BLGF (QScore=0 at BLGF) | 6 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |

**SCORE NODE LIST**

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 20 | 10 | 20 | 10 | 20 | 10 | 20 | 20 | 20 | 20 | 20 |
| 3 | 40 | 20 | 40 | 20 | 40 | 20 | 40 | 40 | 40 | 40 | 40 |
| 4 | 60 | 30 | 60 | 30 | 60 | 30 | 60 | 60 | 60 | 60 | 60 |
| 5 | 80 | 40 | 80 | 40 | * | 40 | * | * | * | * | * |
| 6 | 100 | 50 | 100 | 60 | | 60 | | | | | |
| 7 | | 60 | | 80 | | * | | | | | |
| 8 | | 70 | | 100 | | | | | | | |
| 8 | | 100 | | | | | | | | | |

**PERCENTAGE OF CANDIDATES SCORING ON EACH NODE**

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 62 | 78 | 90 | 90 | 92 | 92 | 92 | 91 | 92 | 92 | 92 |
| 2 | 32 | 10 | 9 | 2 | 7 | 0.2 | 7 | 8 | 7 | 7 | 7 |
| 3 | 5 | 9 | 1 | 7 | 0.9 | 7 | 0.9 | 0.9 | 0.9 | 1 | 1 |
| 4 | 0.5 | 2 | 0.1 | 0.2 | 0.1 | 0.01 | 0.08 | 0.1 | 0.05 | 0.1 | 0.1 |
| 5 | 0.03 | 1 | 0.01 | 1 | * | 1 | * | * | * | * | * |
| 6 | 0.01 | 0.2 | * | 0.07 | | 0.08 | | | | | |
| 7 | | 0.1 | | 0.01 | | * | | | | | |
| 8 | | 0.01 | | * | | | | | | | |
| 8 | | * | | | | | | | | | |

Percentages of candidates less than 0.01 and nodes not appearing in 25000 cycles are marked *

**Table 5 MRQ 5/25 partially constrained selection**

## MSS Results from 'Extended Matching Item' (EMI) and 'Drag and Drop' (DD) Question Types

Extended matching item and drag and drop question types share the same process of selection. In either case the candidate is required to select a number of items from a list then enter or move them to their correct positions. Thus the candidate must make two selections - which item and where to put it. A candidate with zero knowledge will make both decisions randomly and this decision making process is mimicked in the Marking Simulator software by matching random selection of items with a random selection of positions for each iteration.

The detailed scoring for this question type may vary from assessment system to assessment system. In its simplest form, a positive score is allocated for each item correctly positioned. In the TRIADSystem, each item may be allocated a different score for each position and the question may have dummy items with no correct positions and dummy positions that score none of the items. This allows questions to be developed that require the candidate to select the most appropriate item for a position whilst correct but less appropriate items would attract a lower positive score in the same position. Clearly the possibilities for varying the scoring in this question type are legion and it is thus important that the outcomes are investigated using the Marking Simulator before settling on one of the scoring regimes indicated above.

In constrained versions of this question type where all positions must be filled, the effect of general penalty negative scores for inappropriate positioning of items becomes less important because the incorrect positioning of one item is automatically exerts a penalty. This is because the inappropriate item does not score in the position selected and the item that should have scored in that position must now be placed in a position where it cannot score. Thus to apply an element of negative scoring could be regarded as exerting a double penalty. Constrained versions of this question type will show a lack of scoring nodes at high values because of this property.

In partially constrained versions of this question type the effect on the availability of scoring nodes is more limited because the candidate can choose not to position some of the items. This effect is illustrated in Table 6 by the data from a 5x5 EMI question. Scoring nodes at 20% intervals are always available with additional nodes available depending upon the level of negative scoring. Interestingly, for penalty settings of -20%, the partially constrained EMI 5x5 (Table 6) has identical BLGF properties to the partially constrained MRQ 5/25 example (Table 5). In this configuration 91% of guessing candidates score zero and 98% score 20% or less.

**EMI BASE LEVEL GUESS FACTOR (BLGF) PARAMETERS**

| | | |
|---|---|---|
| Number of correct items: 5 | Number of dummy items: 0 | Each item may be selected: x1 |
| Number of correct positions: 5 | Number of dummy positions: 0 | nItems >= nPositions |
| Open selection up to number of positions. | Negative scores resolved to zero | |
| Number of iterations: 25000 | Residual BLGF = (((Q%-BLGF)/(100-BLGF)*100) | |

| | Negative scores on incorrectly positioned items | | | | |
|---|---|---|---|---|---|
| Parameter | 0 | -5 | -10 | -15 | -20 |
| BLGF% | 12 | 7 | 4 | 3 | 2 |
| % passing Q'n at 40% | 12 | 3 | 2 | 2 | 1 |
| % passing Q'n at 40% of (100-BLGF) | 2 | 2 | 1 | 1 | 0 |
| 40% pass-mark equiv't at 100-BLGF | 47 | 44 | 42 | 42 | 41 |
| Residual BLGF (QScore=0 at BLGF) | 7 | 5 | 3 | 3 | 2 |

SCORE NODE LIST

| | | | | | |
|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 |
| 2 | 20 | 5 | 10 | 5 | 20 |
| 3 | 40 | 10 | 20 | 10 | 40 |
| 4 | 60 | 15 | 30 | 20 | 60 |
| 5 | 80 | 20 | 40 | 25 | 80 |
| 6 | 100 | 25 | 50 | 30 | 100 |
| 7 | | 30 | 60 | 40 | |
| 8 | | 35 | 70 | 45 | |
| 9 | | 40 | 80 | 60 | |
| 10 | | 50 | 100 | 65 | |
| 11 | | 55 | | 80 | |
| 12 | | 60 | | 100 | |
| 13 | | 75 | | | |
| 14 | | 80 | | | |
| 15 | | 100 | | | |

PERCENTAGE OF CANDIDATES SCORING ON EACH NODE

| | | | | | |
|---|---|---|---|---|---|
| 1 | 54 | 62 | 78 | 82 | 91 |
| 2 | 35 | 8 | 10 | 6 | 7 |
| 3 | 10 | 8 | 7 | 3 | 1 |
| 4 | 1 | 6 | 2 | 4 | 0.2 |
| 5 | 0.2 | 4 | 2 | 2 | 0.05 |
| 6 | 0.01 | 4 | 0.6 | 0.9 | 0.01 |
| 7 | | 3 | 0.2 | 0.9 | |
| 8 | | 2 | 0.1 | 0.5 | |
| 9 | | 0.9 | 0.02 | 0.2 | |
| 10 | | 0.9 | 0.01 | 0.1 | |
| 11 | | 0.6 | | 0.04 | |
| 12 | | 0.1 | | 0.01 | |
| 13 | | 0.08 | | | |
| 14 | | 0.04 | | | |
| 15 | | 0.01 | | | |

**Table 6 EMI 5x5 partially constrained selection**

In the constrained configuration (data not tabulated) with a zero penalty, the 80% node is lost and the BLGF is around 21% compared to the 12% value for the partially constrained version. With a 20% penalty, the constrained BLGF scores are similar to those of the partially constrained variant but there are no scoring nodes between 21% and 99%. This means that the constrained question does not discriminate between partly correct answers as well as the partially constrained variant does. In practice, it is likely that a high proportion of the candidates will be tempted to position all items when the number of items is small and the same as the number of positions. The BLGF properties of a partially constrained question will then approach the properties of the constrained version with the loss of higher value scoring nodes. It may be advisable therefore to increase the number of positions and include dummy items in order to minimise this effect.

There are many other issues and experiments that could be discussed with respect to EMI questions. Space precludes further discussion here but the effect of the inclusion of dummy items, dummy positions and of positions that can score two or more items differently are currently being investigated and will be reported at a later date.

## Discussion and Application of the Results

The Marking Simulator is designed to inform tutors of the likely score distributions resulting from a range of question types and scoring regimes in order to facilitate better question design. The most critical feature of the question design is to ensure that the academic content of the question is appropriate and unambiguous. The candidate should not easily be able to eliminate options in order to reduce the number of possibilities for selection.

Having satisfied the academic conditions, it remains to allocate a scoring regime to the question that is appropriate to the purpose of the assessment. In a competency-based assessment, where maybe only a pass/fail decision is required for each question then the BLGF parameters may be used to allocate an appropriate pass-mark for each question to produce a zero(fail) or 1(pass) output for the question while at the same time minimising the chance that a candidate may achieve a pass by guesswork alone.

In norm-based assessments, where it may be desirable to grade candidates within questions as well as across the whole assessment, a more detailed examination of the BLGF parameters may be desirable. This is to ensure maximum discrimination between candidates after filtering out as higher proportion of the scores attainable by guesswork alone as possible. In such instances, it is advisable to select a scoring regime that reduces the BLGF to a minimum whilst retaining as many evenly distributed scoring nodes as possible. Ideally, the higher scoring nodes should have the lowest possible frequency of candidates scoring by guesswork. Care must be taken to ensure that the higher value nodes do not merely represent increased numbers of errors and that a candidate making less errors does not achieve a lower score as seen in the 2/5 MRQ examples. This will ensure that high scoring nodes are available for candidates with knowledge/understanding of the correct answer to be scored and graded according to their ability whilst high scores

are difficult to achieve by guesswork alone. Highly discriminating questions have the potential to generate more detailed feedback to tutors on the degree to which learning has been achieved.

In some cases however it may be necessary on academic grounds to include questions that are not ideal in terms of scoring because their structure imparts a high BLGF when compared to the overall pass-mark for the test. In such cases it may be necessary to correct the scoring by reference to the BLGF data in order to output a mark distribution that shows the highest discrimination and reflects the range of abilities in the cohort.

Various BLGF compensation strategies are and a study of the way in which scoring node distribution may affect item statistics and of what are the most valid correction factors is ongoing so that statistically sound recommendations can be made to tutors.

## Conclusions

The preliminary investigations outlined in this paper have indicated that there are some scoring configurations that can have unexpected outcomes for the unwary tutor, even for apparently simple question styles. The outcome of more complex scoring schemes is even more difficult to predict intuitively. Examples of these might include Extended Matching Item question types where there may be a range of answers possible for some positions, each carrying a different score. In these situations, the Marking Simulator can be invaluable in helping the tutor to develop high quality questions with a scoring methodology that produces a low mathematical guess factor (BLGF) and generates results that are both discriminating and academically sound.

The preliminary results outlined here have encouraged us to continue developing the Marking Simulator with the addition of a user-friendly interface, graphical output and an extended range of question types. Although results have been reported in tabulated form here, the final Marking Simulator software will additionally report results in graphical format to aid ease of interpretation by tutors.

It is to be hoped that the informed use of the Marking Simulator will be a positive step towards increasing the quality of computer-based assessments. However, it is worth noting here that the output from the Marking Simulator should be seen merely as an aid to good question-structure design. Question content, together with gender, psychological and previous test experience of the candidate may also have substantial effects on the scoring outcomes for an assessment.

## References

Boyle,A., Hutchison,D., O'Hare,D. & Patterson,A. (2002) Item selection and application in Higher Education. Proceedings of the 6th Annual CAA Conference. (*This volume*)

Burton, R & Miller, D. (2000) Why tests are not as simple as a,b or c. The Times Higher Education Supplement, February 4th. P42

Ebel, R.L. & Frisbie, D.A. (1991) *Essentials of educational measurement*. 5th Edition Prentice-Hall

Huff, K.L. & Sireci, S.G. (2001) Validity Issues in Computer-Based Testing. Educational Measurement: Issues and Practice. 20, 3, 16-25

Mackenzie, D.M. (1999) Recent Developments in the Tripartite Interactive Assessment Delivery System (TRIADS). Proceedings of the 3rd Annual CAA Conference ISBN 0953321037

Rust, J & Golombok, S. (1999*) Modern psychometrics: The science of psychological assessment*. Part One, 45-47. Taylor & Francis

Ryle, A.P. (1996) Objective Tests: In Defence of 'Negative Marking'. Life Sciences Educational Computing. Vol 7, No 1.