ESRC National Centre for Research Methods

Use of the LEMMA Online Learning Materials

Jo-Anne Baird, Rebecca Pillinger and Fiona Steele

Report prepared for the LEMMA (Learning Environment for Multilevel Modelling and Applications) node, University of Bristol

January 2012

Acknowledgements

The LEMMA project is funded under the ESRC National Centre for Research Methods (NCRM). The LEMMA online course was established in April 2008 under Phase 1 of NCRM (grant number RES-576-25-5006) and further developed under Phase 2 (RES-576-25-0003). A third three-year project began 1 October 2011 (RES-576-25-0032).

The LEMMA online course

The LEMMA (Learning Environment in Multilevel Modelling and Applications) online training system has been developed by the Centre for Multilevel Modelling, with the first release available in April 2008. The course was developed under the LEMMA research project, a node of the ESRC-funded National Centre for Research Methods. The LEMMA course is housed in Moodle. After registration, the materials can be accessed free of charge. An overview of the materials and registration details can be found at http://www.bris.ac.uk/cmm/learning/course.html.

The primary target audience for the course is researchers from the social sciences and public health, at all career stages and from both academic and non-academic sectors. With this audience in mind, the materials use examples from a range of social science disciplines and public health.

An important aim of the course was to cater for all levels of learner, from those with minimal experience of traditional elementary statistical methods (possibly needing a refresher) to more advanced users, including those who wish to train others in multilevel modelling. The course was therefore designed so that learners can enter at different points according to their prior experience of statistical modelling.

Most modules consist of a 'concepts' and 'practice' component. The 'concepts' gives a detailed description of a statistical model, including the types of data and research questions it can be used to investigate and its interpretation, but without reference to any statistical software package. The 'practice' component then provides instructions for the analysis of a particular dataset using a range of software packages, currently MLwiN, Stata and R.

Overview

The 'Learning Environment in Multilevel Modelling and Applications' (LEMMA) online learning modules were initiated in April 2008, with additional modules and materials being added subsequently. The environment currently contains nine course modules, including practical materials in MLwiN, Stata and R. More modules are at the planning stage.

The materials have been very popular internationally. Data for this report were extracted in September 2011 when there were already over 6,000 registered users from over 100 countries. Over half of the users were postgraduate students and nearly 20% were academics. As anticipated, the most frequent primary discipline of registered users fell under the social sciences category (nearly 60%). Nonetheless, users were also researchers in the public sector, charities and private organisations and a large number of disciplines were represented. This demonstrates the need and demand for advanced, online statistics training.

Materials were presented as 'concepts' learning packages, together with practical exercises in a range of statistical software packages and quizzes devised for self-assessment for many of the modules. Production of the quiz items is time-consuming, yet they were the least frequently used aspect of the materials. However, online users will only receive feedback regarding their understanding of the concepts from the quiz items, in the absence of any face-to-face tuition. Therefore, usage of the quiz items will be monitored before a decision is taken about whether to discontinue their production.

Users' performances on the quiz items was analysed and some of the findings are presented here. Causes of performance are confounded, as item difficulty is a product of format, language used, position in the quiz, demand of the concepts and so on. In a number of cases, it was possible to hypothesise the probable cause of difficulty and this led to a revision of the item presentation. After all, the purpose of the quiz is to assess the key concepts without the presentation obstructing responses as far as possible.

Unlike in other published literature, two of the course tutors were able to judge the order of difficulty of items to a significant (although moderate) extent. This is likely due to the control of curriculum materials in this case.

1. Introduction

1.1 The LEMMA online course

The LEMMA (Learning Environment in Multilevel Modelling and Applications) online training system has been developed by the Centre for Multilevel Modelling, with the first release available in April 2008. The course was developed under the LEMMA research project, a node of the ESRC-funded National Centre for Research Methods.[1] The LEMMA course is housed in Moodle, a free web application for building e-learning systems. The materials can be accessed free of charge, although users are required to register first. An overview of the materials and registration details can be found at http://www.bris.ac.uk/cmm/learning/course.html.

There were several guiding principles in the design of the course:

(i) The materials should be accessible to anyone with a basic statistics training;

(ii) The materials should be consistent (in terms of notation, conceptual framework and layout) and carefully sequenced to allow learners to progress to advanced quantitative methods;

(iii) Practical exercises should be provided in more than one statistics package;

(iv) Although the course is unsupported, quizzes are provided to allow learners to evaluate their understanding of the material;

(v) To monitor usage and inform future training initiatives, the web platform should allow collection of basic data on the profile of learners (e.g. their employment sector, academic discipline and prior exposure to statistical methods) and usage of the materials.

The primary target audience for the course is researchers from the social sciences and public health, at all career stages (from undergraduate students to experienced researchers) and from both academic and non-academic sectors. With this audience in mind, the materials use examples from a range of social science disciplines and public health.

An important aim of the course was to cater for all levels of learner, from those with minimal experience of traditional elementary statistical methods (possibly needing a refresher) to more advanced users, including those who wish to train others in multilevel modelling. The course was therefore designed so that learners can enter at different points according to their prior experience of statistical modelling. Consistent terminology and notation, and cross-referencing with earlier modules, was used to facilitate progression from introductory to advanced modules.

An important feature of module design is the separation of concepts and practice. The 'concepts' component of a module gives a detailed description of a statistical model, including the types of data and research questions it can be used to investigate and its interpretation, but without reference to any statistical software package. The practical component then provides instructions for the analysis

[1] Two three-year phases of LEMMA have been funded from 2005 (grant numbers RES-576-25-5006 and RES-576-25-0003). A third three-year project began 1 October 2011 (RES-576-25-0032).

of a particular dataset. The aim of this design is to separate learning statistical concepts and learning how to use a software package. Another advantage of separating out concepts and practice is that the materials can be easily extended to include practical exercises using different software packages and datasets from different subject areas. The LEMMA course currently contains practical materials in MLwiN, Stata and R.

This report is based on data collected up to the end of September 2011 when the course contained the following eight modules:

1. Using quantitative data in research
2. Introduction to quantitative data analysis
3. Multiple regression (including practicals in MLwiN, Stata and R)
4. Multilevel structures & classifications
5. Introduction to multilevel modelling (including practicals in MLwiN, Stata and R, and an extended application to The Use of Performance Indicators in Education)
6. Regression Models for Binary Responses (including practicals in MLwiN, Stata and R)
7. Multilevel Models for Binary Responses (including practicals in MLwiN, Stata and R)
8. Multilevel Modelling in Practice: Research Questions, Data Preparation and Analysis

Further modules are in development on the following topics: (i) three-level models; (ii) cross-classified multilevel models; (iii) multiple-membership multilevel models; and (iv) single-level and multilevel models for nominal responses. Modules on longitudinal data analysis are also planned.

The online format has disadvantages in terms of reduced and de-personalised interaction with and between learners, but it also has significant benefits. Being online, the LEMMA materials are more flexible and widen access to students due to reduced cost, in terms of time and money, compared to attending a face-to-face workshop. Learners can also access the materials at a time and pace that is suitable for them, so scheduling issues with face-to-face workshops and their own research programme are less problematical. The online format might be more attractive to non-academics than face-to-face workshops due to the flexibility and difficulty in dedicating several days to a workshop.

1.2 Structure of the report

This report outlines the nature of the LEMMA course users, including their statistical experience. It goes on to look at patterns of use, including use of quizzes and the performance of students who have tackled the quizzes. Some of the quiz responses indicate issues about the design of the items, whilst others appear to show which concepts were difficult for users. These are discussed in this report, but readers are also referred to a report on a qualitative analysis of the design of the items (Ahmed, 2011). Finally, we investigate the relationship between trainers' perceptions of quiz item demands and performance on the items. Figures for this report were produced on 30 September 2011.

2. LEMMA users

Registered users of LEMMA (*n*=6,076) are predominantly European (54.0%) and many are from the UK (31.1%). Many users are from North America (23.9%), but there were users from over 100 countries, in every continent. **Table 1** shows the countries with the highest number of users. Assuming 30 learners in each training session, face-to-face workshops would have to be run 202 times to reach this number of individuals. Although the online materials can also be used to reinforce and refresh learning from face-to-face workshops, only a small proportion of learners had attended one of the Centre for Multilevel Modelling's workshops (8%).

Table 1 Countries with at least 20 LEMMA users

Country	Number of users	Percent of total
United Kingdom	1888	31.1
United States	1220	20.1
Netherlands	262	4.3
Germany	250	4.1
Australia	243	4.0
Canada	233	3.8
Italy	161	2.6
Spain	141	2.3
Belgium	131	2.2
Switzerland	109	1.8
Sweden	95	1.6
India	94	1.5
France	88	1.4
Brazil	64	1.1
Korea, Republic of	61	1.0
South Africa	59	1.0
Ireland	54	0.9
China	51	0.8
Norway	48	0.8
Finland	36	0.6
Kenya	35	0.6
Japan	34	0.6
Mexico	34	0.6
Hong Kong	31	0.5
Taiwan	31	0.5
Chile	30	0.5
New Zealand	29	0.5
Austria	27	0.4
Colombia	26	0.4
Portugal	25	0.4
Greece	22	0.4
Turkey	21	0.3
Denmark	20	0.3
Other countries	423	7.0
Total	6076	100

Whilst there is an aim to build capacity among both academic and non-academic researchers, it was anticipated that the largest uptake would be amongst postgraduate students and this was indeed the case (48.3%), with doctoral students being the most prevalent (31.4%). Most users were

university students (51.1%) and the second largest group were academics (17.4%). Over 1000 people who were employed in the public sector (11.5%), by a private research institute (2.6%), industry (1.8%) or by a charity (1.0%), used the LEMMA materials. Most people registered on the VLE for self-learning purposes (87.2%). Some courses had recommended the site and this drove some registrations (4.3%). The remaining registrations were from people who were attending a face-to-face workshop run by the Centre for Multilevel Modelling (3.1%) or who were considering using the materials in their own teaching (5.3%).

Over 20 disciplines were represented amongst users' primary disciplinary backgrounds, with statistics, medical sciences, and psychology being the most prevalent (**Table 2**). The materials were written with social scientists as the target audience and most of the examples are from social science. Around 60% of users were from the social sciences. A broad range of disciplines was represented, some of which do not have a strong quantitative tradition (e.g. Social Anthropology, Socio-Legal Studies, and Arts and Humanities disciplines). Nearly half of the learners already had a Masters and over one third had a doctorate qualification, with only 4% not having at least a degree.

Table 2 Primary discipline of LEMMA users

Discipline	Number of users	Percent
Statistics, methods & computing	900	14.9
Medical sciences	880	14.6
Psychology	855	14.2
Sociology	530	8.8
Economics	444	7.4
Education	439	7.3
Management & business studies	275	4.6
Political science & international studies	261	4.3
Biological sciences	216	3.6
Human geography	183	3.0
Demography	153	2.5
Social policy	115	1.9
Environmental science	67	1.1
Science & technology studies	64	1.1
Social work	53	0.9
Engineering & physical sciences	48	0.8
Linguistics	38	0.6
Area studies	31	0.5
Socio-legal studies	23	0.4
Environmental planning	22	0.4
Economic & social history	19	0.3
Arts & humanities	18	0.3
Social anthropology	16	0.3
Other	381	6.3
Total	6031	100

3. Users' previous experience in statistics

Overall, most learners were at least reasonably experienced in using statistical techniques (**Table 3**), but some learners had little or no practical experience.

Table 3 Learner overall statistical experience

	Number of users	Percent
No practical experience	344	6.1
Some experience, e.g. for assessed coursework	1402	24.7
Reasonably experienced (sometimes use in job)	2267	39.9
Very experienced (regularly use in job)	1666	29.3
Total	5679	100

The LEMMA curriculum was designed to allow an entry point well below the level of multilevel modelling, to provide a progression ladder for learners. Our observation from the face-to-face workshops has been that learners are sometimes not equipped to deal with the multilevel materials because they had no experience of statistics at a more basic level, or that their experience had been too long ago. Most learners had used descriptive statistics, t-tests, analysis of variance, simple, multiple and logistic regression. The first LEMMA module begins with basic ideas of quantitative research, the second builds upon this to look at quantitative analysis and the third moves on to regression. Learners are introduced to multilevel data structures in module 4, followed by multilevel modelling in module 5. Approximately 7% of learners were not familiar with simple regression and 19% had received training in it, but had never used it in their own research (see Figure 1). This demonstrates the range of learners for whom the LEMMA materials had to cater. Future developers of online materials could also consider the strategy of creating a progression route so that people who do not have the pre-requisites can gain them.

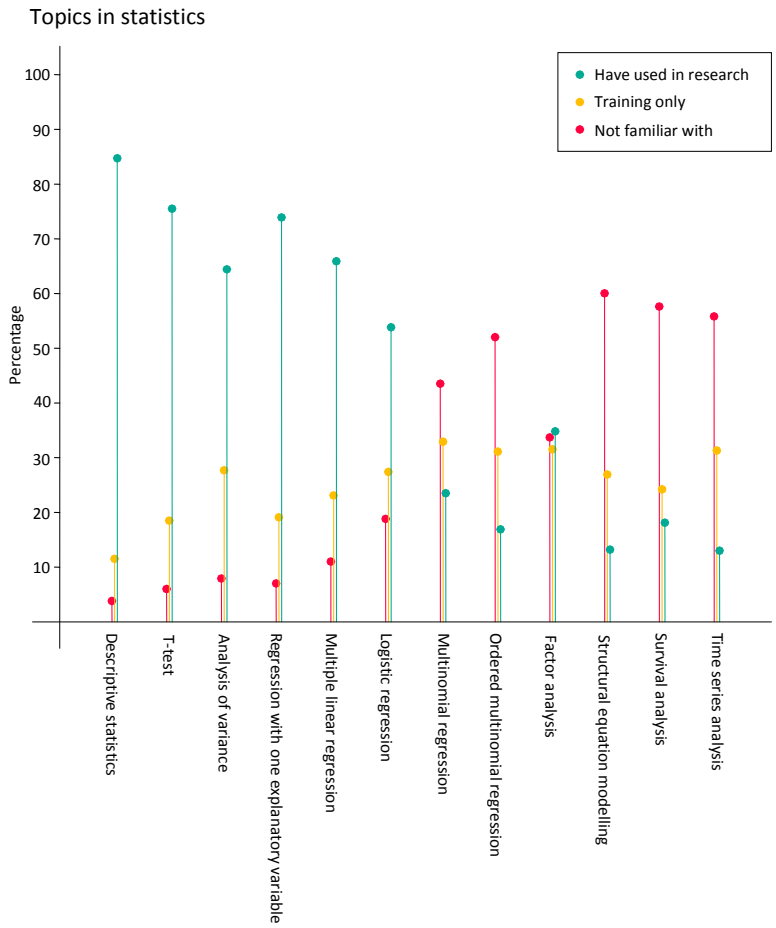


Figure 1 User familiarity with statistical topics

Decisions had to be taken about the extent to which the course materials should assume mathematical expertise and present mathematical details of methods (e.g. proofs and descriptions of estimation procedures). As social scientists are the main target audience, it was decided to assume only a minimal level of mathematics, probability theory and statistics, and users' ratings of familiarity with mathematical and statistical topics suggest that this was the correct course of action (Figure 2). Only 23% of users were confident in applying matrix algebra, for example, so had the course content laid out the concepts using matrix notation, it would have been inaccessible to most users. As far as possible, more advanced topics (such as Markov chain Monte Carlo estimation in Module 7) were explained in an intuitive way, with the emphasis placed on practical issues for applied researchers. Nevertheless, as previously discussed, the LEMMA materials also attracted advanced users (14.9% are statisticians and 7.4% are economists) and to cater for those, Technical Appendices are provided for Modules 6 and 7 on single-level and multilevel models for binary responses.

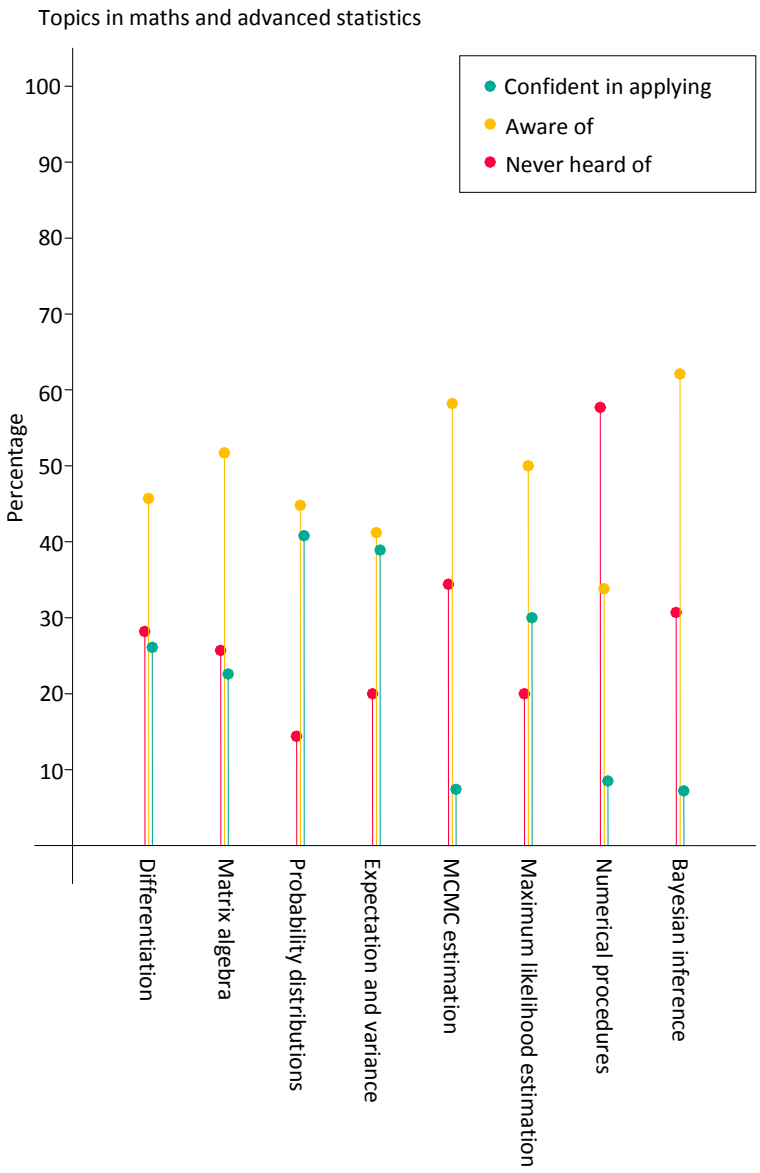


Figure 2 User familiarity with mathematical and advanced statistics topics

First-time users of LEMMA are strongly encouraged to take the 'pre-requisites' quiz, which aims to help learners to assess their familiarity with introductory-level statistical concepts. Learners who find this quiz difficult are advised to study Modules 1 and 2 to refresh their knowledge. Just under half of all registered users attempted at least one question on the pre-requisites quiz (*n*=2,813), with approximately 80% of those completing the quiz (*n*=2,108). Scores on this quiz give an indication of learners' preparedness for the course and in general those who attempted the quiz did well (**Table 4**). Ahmed (2011) outlined the language barriers that could have been contributory factors to the low success rates on PR 04(ii), PR 07 and PR 12. Additionally, question PR 04(ii) is a reminder (through the feedback) to the learner that nonlinear relationships could exist in the data, even if there is no correlation. In the context of the question, it is easy to slip into less precise thinking and

answer the question wrongly. Questions PR 07 and PR 12 relate to confidence intervals, which are well known to be conceptually slippery. delMas (2011, p89) wrote that

> "… the meaning of a 95% confidence interval is based on the understanding that there is a 95% chance that a single randomly selected sample will be one of the samples that provides a confidence interval that captures the population characteristic. This understanding requires a complex mental model of several related concepts, which alone may make reasoning from confidence intervals difficult."

As such, it is to be expected that these questions might cause difficulty for learners and the feedback that they received should have assisted their understanding of the concepts. Of course, we have general feedback about the LEMMA materials from learners and some have sent messages about specific issues, but we do not know empirically the extent to which the feedback from the quiz items was helpful in improving learning.

Table 4 Attempts and successes on the pre-requisite quiz questions

Question	Number of users	% with correct answer
PR 01	2813	94.8
PR 02	2751	84.0
PR 03(i)	2573	86.0
PR 03(ii)	2480	67.9
PR 03(iii)	2383	92.3
PR 03(iv)	2356	85.0
PR 03(v)	2318	68.9
PR 04(i)	2318	94.0
PR 04(ii)	2313	16.4
PR 04(iii)	2303	78.3
PR 04(iv)	2282	85.1
PR 05	2278	66.9
PR 06	2247	47.9
PR 07	2227	21.4
PR 08(i)	2221	80.9
PR 08(ii)	2219	58.9
PR 08(iii)	2209	80.9
PR 08(iv)	2204	62.6
PR 09	2196	85.5
PR 10	2178	60.7
PR 11	2174	69.3
PR 12	2170	22.9

The correlation between learners' self-rating of statistical experience and their performance on the pre-requisite quiz was low (r=0.36). Thus, it might be important to include pre-requisite 'reminder' curricula and tests when designing online materials (the rationale for Modules 1-3 in LEMMA). Users might consider themselves experienced statistically, but have forgotten (or never known) some of the fundamental concepts upon which a target curriculum builds to more complex ideas.

4. Patterns of use

The materials were designed to permit on-screen usage, as well as off-line use of printed pdfs. To facilitate this, the curriculum was presented in short lessons and materials were separated into concepts and practicals. As we wished learners to be able to view the materials on-screen or in printed form, we did not use many of the affordances available in electronic media. A large minority (38%) of users availed themselves of both the pdf and web-based modes of presentation, whilst 58% were web-only users. A very small proportion of learners only used pdfs (4%). Although there is no doubt that computerised methods of presentation have advantages, being able to print materials offered LEMMA course users flexibility. Few users printed more than four pdfs. We would have predicted that the practicals would be printed more frequently than the concepts, as learners could then work through the printed material alongside their on-screen software analyses. However, the concepts materials were just as likely to be printed as the practicals.

Figure 3 shows the proportion of users for whom each module was the entry point into the materials, by content type. Over half of those who looked at concepts materials began at the first module, on introductory concepts related to using quantitative data in research, but a number of users began at Modules 4 or 5, which were the first to introduce multilevel data structures and analysis. This could imply two sets of users, with the first needing to refresh their understanding of basic concepts and the second being confident to begin with multilevel modelling. Module 1 was also the most frequent starting point for those using the quizzes, but Module 3 (multiple regression) was the most frequent starting point for practicals because this was the first one available (Modules 1, 2 and 4 had no practical).

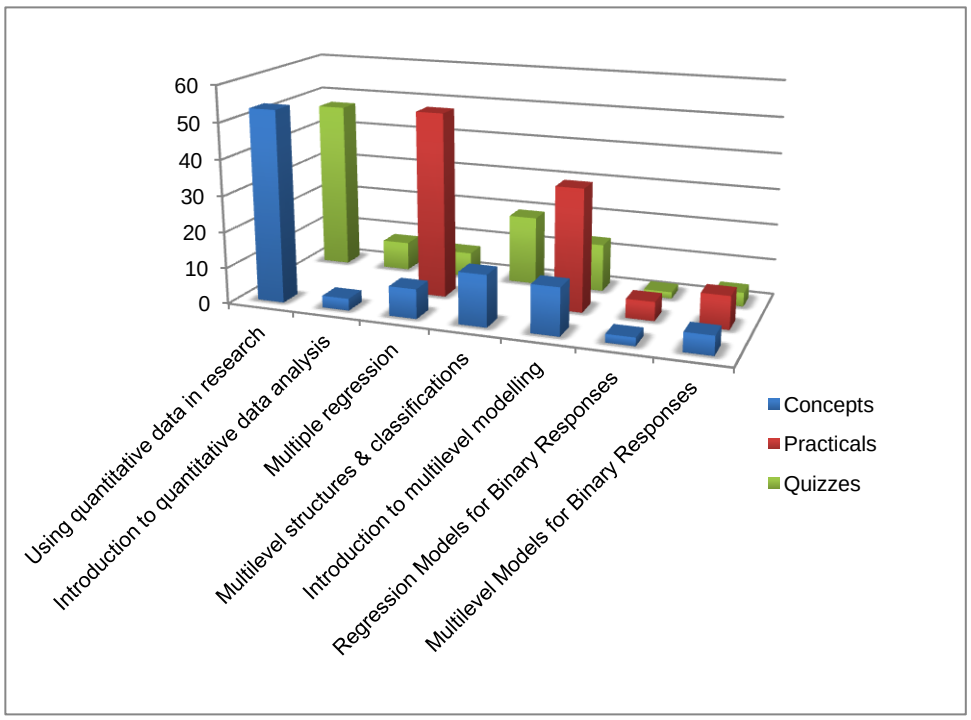


Figure 3 First module accessed, by content type

5. Quiz usage and scores

Writing good quality quiz items and feedback, reviewing them and posting them on the system is resource intensive. Thus, it is important to know whether learners used them. As discussed above, less than half of the registered users attempted the pre-requisites quiz. That might be expected, as a large section of users would be confident enough to use the materials without checking that they had the pre-requisites. However, the numbers using subsequent quizzes gets increasingly smaller through the modules, with the numbers in double digits by Module 6.

As learners were expected to have studied the curriculum material prior to attempting the quiz, we might expect a high success rate on the quizzes. However, many users were new to the material, so their attempts at the quiz can be seen as part of the learning process. As such, it is difficult to interpret the scores. Mean scores for the quizzes were between 42% and 82%, with the median score being 64%. Questions with the lowest success rates involved conducting multilevel analysis and giving values or specifying and estimating a multilevel model and supplying a numerical response based on the results (Module 5, Sections 5.4 and 5.5). This open-response format reduces the likelihood of selecting a correct answer by chance.

Of the 144 quiz items (excluding the 22 pre-requisite quiz items), approximately one in 10 had success rates of less than 30%. The following takes a few of the items with a lower than 30% success rate to investigate possible causes. The first of these occurs in Module 1, Section 1.4, dealing with data hierarchies. The success rate on this item was only 19% (*n*=768). Users typically gave an answer that involved a multilevel structure (pupils at level 1 and schools at level 2) and therefore did not select the correct answer (schools only, see **Figure 4**). There is no reason to change this item, as it serves as a useful reminder that not all research questions involve multilevel structures.

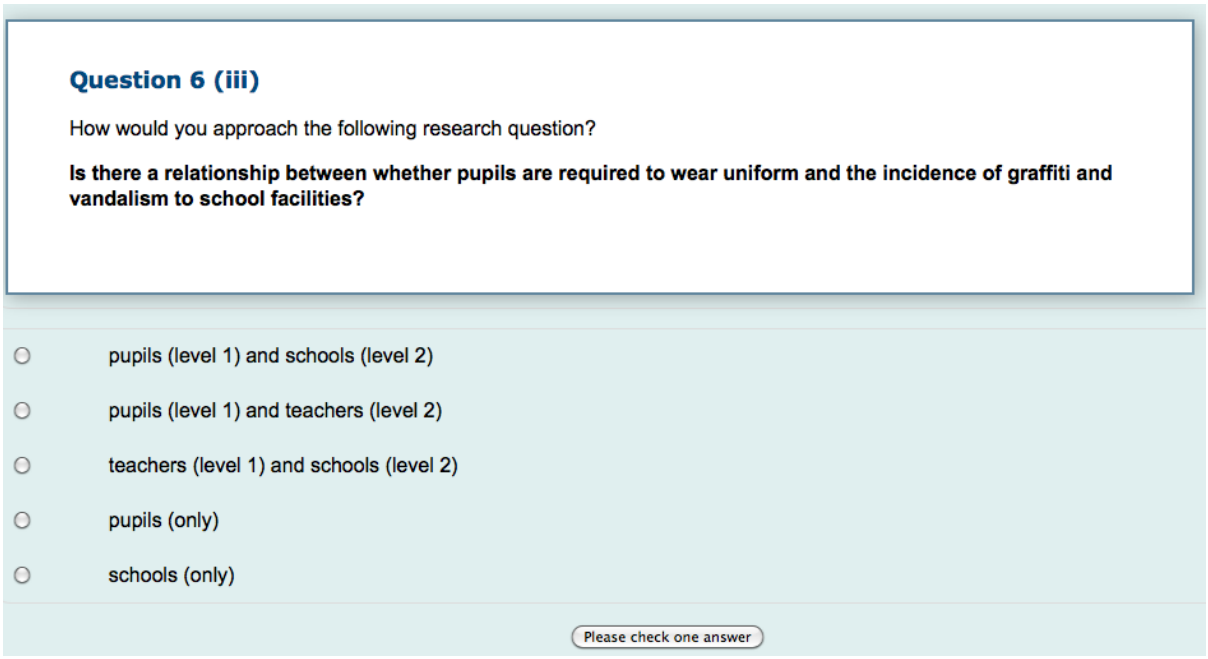


Figure 4 Item 6(iii) in Module 1, Section 1.4

Item 2(iii) in Module 2, Section 2.1 (**Figure 5**) involved using an equation from the previous item and it is likely that this produced the low success rate of 12% (*n*=556). Most users did not give a response to this question (84.2%). An alternative would be to give the equation with the question, however if the educational point involves having learners select the correct equation, questions of this nature are entirely appropriate.

Question 2 (iii)

Again here is the table giving summary statistics for an exam taken by students at three different schools. For schools A and B, the exam was marked out of 100. For school C, the exam was marked out of 1000.

Table 2

	School A	School B	School C
Number of pupils taking exam	**201**	**251**	**301**
Mean exam score	50	55	450
Median exam score	55	55	400

For each school, each pupil's exam score is subtracted from the mean exam score for the school, these differences are squared, and then the squared differences are added up. In other words, if x_i are the exam scores and $\bar{x}$ is the mean of the exam scores for the school, then we take

$\sum(\bar{x}-x_i)^2$. The results of this calculation are shown below:

Table 3

School A	School B	School C
57,800	55,292	2,550,791

Calculate (to 1 decimal place) the appropriate measure for School A, to compare the spread of exam scores across the schools.

Figure 5 Item 2(iii) in Module 2, Section 2.1

Item 4(iii) in Module 2, Section 2.1 involved selecting the appropriate histogram for a set of data (**Figure 6**). The low success rate of 24% (*n*=483) might have been because histograms are typically presented for equivalent ranges and therefore the height of the bars, rather than the area are the pertinent aspects to attend to. In the concepts section, the example presented involved equivalent ranges on the x axis and it might be helpful to amend that to show an example with different ranges, so that learners become familiar with the concept more deeply before tackling this item.

Question 4 (iii)

The table below shows the number of people from various age groups who responded to a survey.

Table 5

Age range	Number of respondents	%	Cumulative %
18 – 25	84	11.8	11.8
26 – 30	27	3.8	15.6
31 – 35	42	**A**	21.5
36 – 50	120	16.9	38.4
51 – 65	213	30.0	**B**
66 – 100	224	31.5	100.0
Total	710	100.0	100.0

Which of these graphs is a histogram of this data?

Figure 6 Item 4(iii) in Module 2, Section 2.1

Item scoring has probably contributed to the low success rate (29.1%, *n*=323) of the item shown in **Figure 7**, as learners would have to get all seven answers correct to score correctly. Here, the concepts themselves are not difficult, but it would be easy to select the wrong option by accident for at least one out of seven of the responses and therefore get the question wrong.

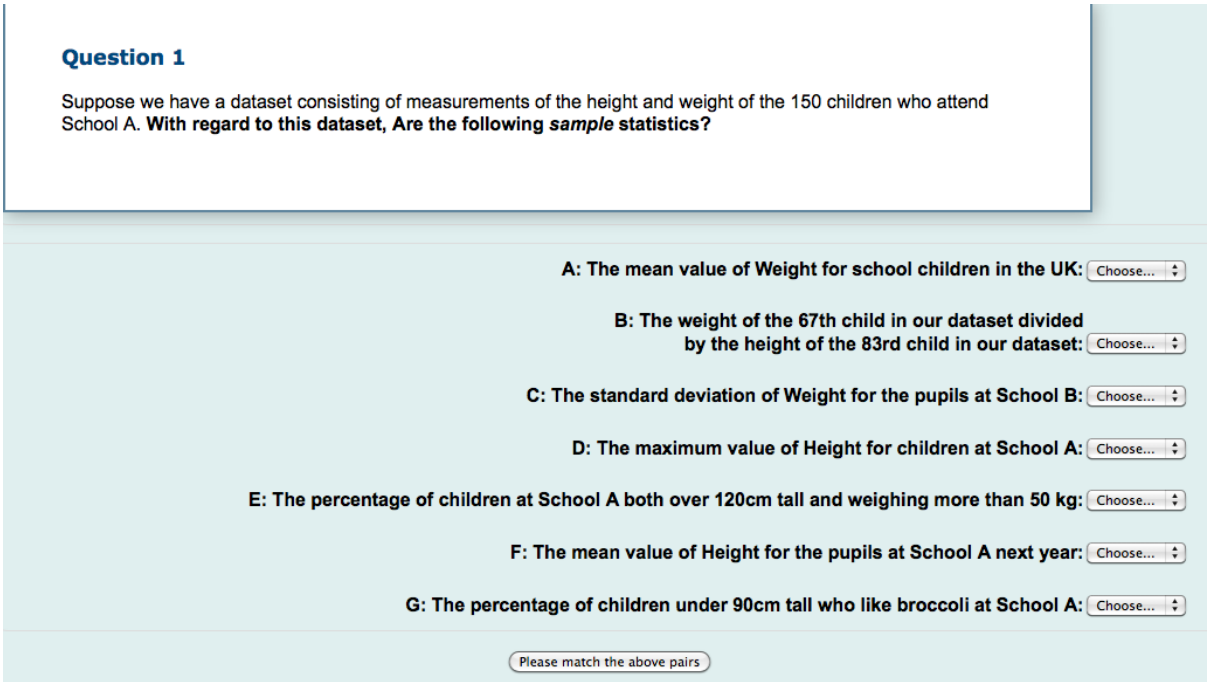
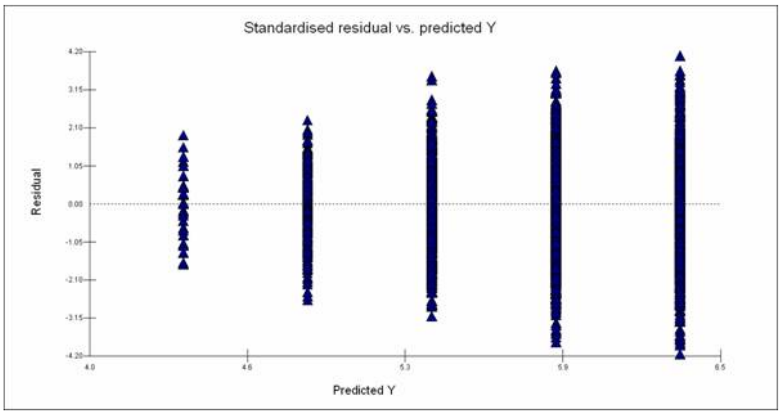


Figure 7 Item 1 in Module 2, Section 2.5

Very few users got the question in **Figure 8** correct (2.6%, *n*=235). Option 1 is designated as the correct answer. However, whilst there clearly are outliers among these residuals, with standardised values of around 4, they do not appear to deviate greatly from the other data at those points on the y axis. Thus, learners might have seen them as problems due to the ill-fit of the model, rather than as deviant data points *per se* and the most frequent response was 'The residuals are hetroscedastic'. Again, partial credit would be a better way to assess this item, with some credit being given for getting a part of the answer correct.

Question 4

What can we infer from the following graph?



Please select only one interpretation.

- There appears to be several outliers, and the residuals are heteroskedastic
- The *y*-values have a non-normal distribution, and the residuals are heteroskedastic
- The *y*-values have a non-normal distribution, and the residuals are homoskedastic
- The residuals are heteroskedastic
- The *y*-values have a non-normal distribution
- There appear to be several outliers
- The residuals are homoskedastic
- none of these

Please check one answer

Figure 8 Item 4 in Module 3, Section 3.5

Only 29.1% (*n*=213) got the item in **Figure 9** correct. The most frequent response was that the value of σ_{u0}^2 would decrease. This was most likely a simple mathematical error, with learners thinking (wrongly) that x-5 on the x axis would correspond with the value of -5 instead of +5.

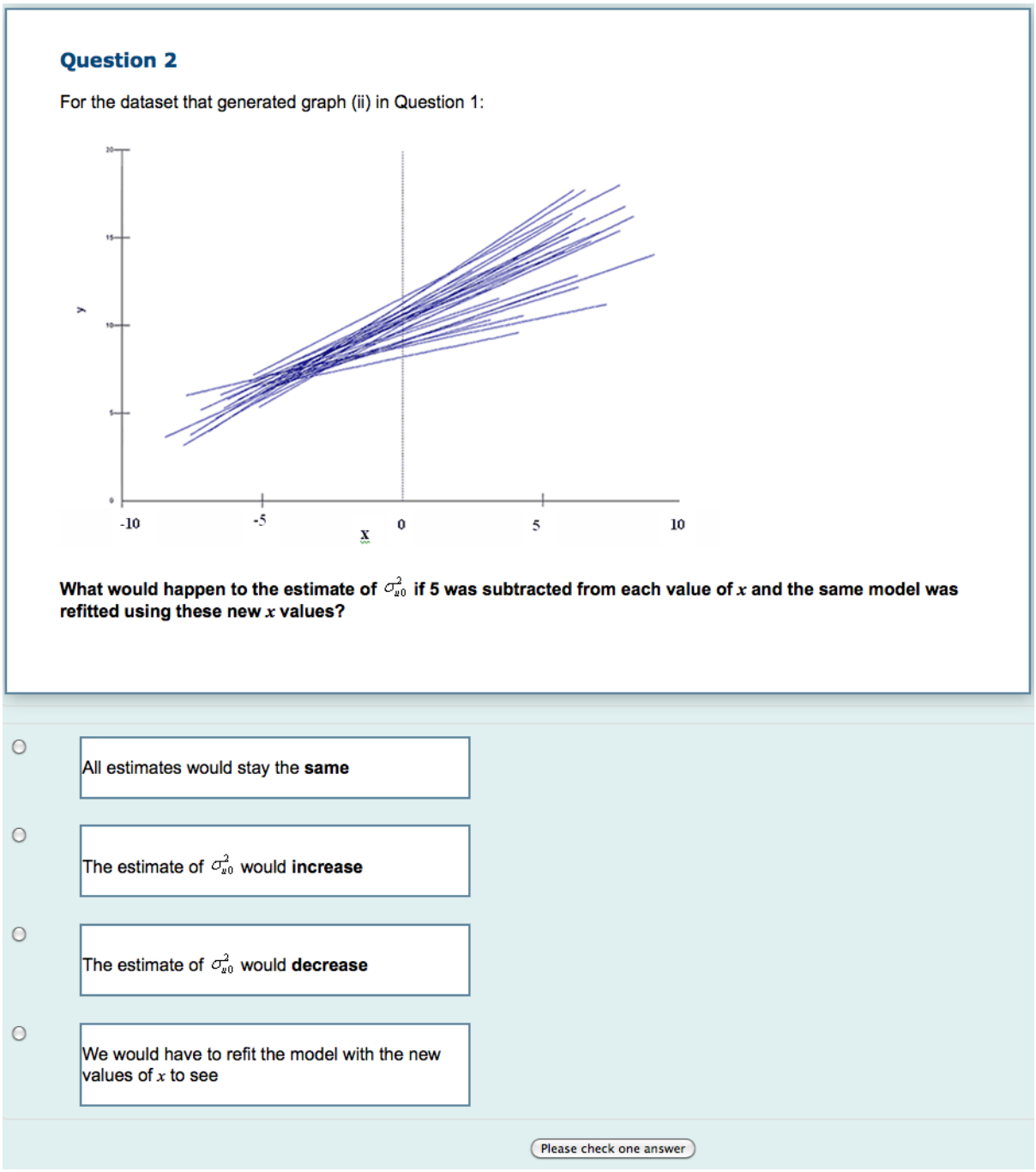


Figure 9 Item 2 in Module 5, Section 5.3

Many people find logistic regression conceptually difficult and this might be one reason why the success rate on the question shown in **Figure 10** was so low (22.6%, *n*=62). However, the question hinges on the word 'change', which the feedback indicates implies an additive rather than

multiplicative effect of changes of x upon y. The wording of this question could be clearer and this might improve the accuracy of responding.

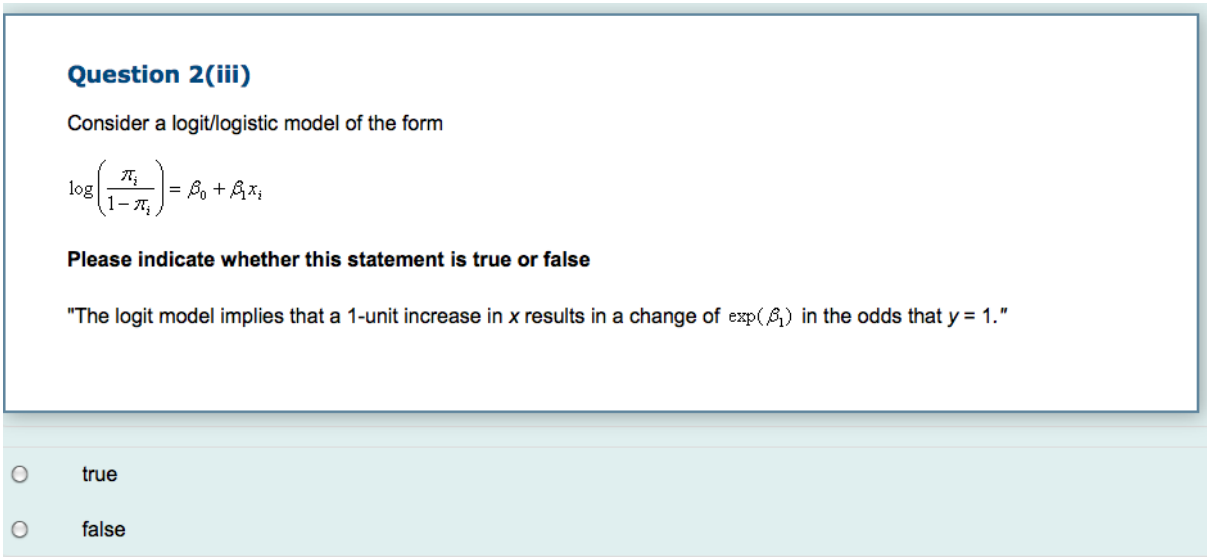


Figure 10 Item 2(iii) in Module 6, Section 6.3

If difficult quiz items put users off from continuing, we would expect a negative association between attrition and success rate. However, no such association was found (**Figure 11**), suggesting that users were willing to persist in the face of failure. Feedback in the form of explanations that were supplied in the event of wrong responses might have been just as useful to learners as knowing that they had correctly understood a concept.

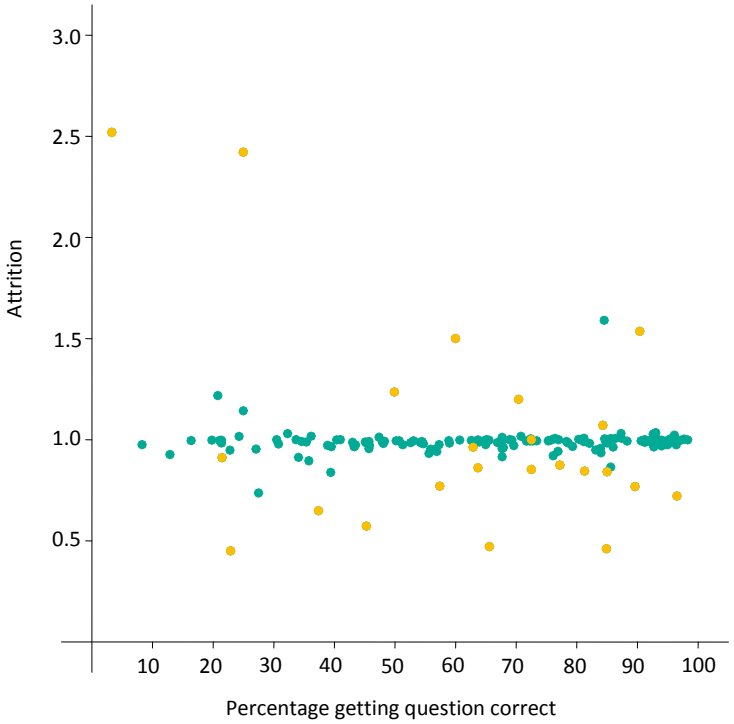


Figure 11 Attrition and success rates on quiz items (ratio of the number attempting each question to the number attempting the previous question plotted against the percentage getting the question correct). Yellow dots represent the first question of each quiz (for which the attrition value may not be so appropriate) and green dots represent the other questions.

6. Judging the level of difficulty of the statistics quiz items

One of the overarching questions for the LEMMA teaching evaluation has been to investigate which aspects of multilevel modelling people find most difficult to learn. Analysis of the quiz questions can illuminate this issue, but we need to be cautious in our interpretation of the statistics on success rates because they are caused by a number of factors in addition to the difficulty of the underlying concept. Researchers have repeatedly demonstrated that adapting the language in which a question is presented can have a large effect upon success rates (Rea-Dickins et al, 2009; Erduran, 2010). Equally, the presentation format (Ahmed and Pollitt, 2010) can have an impact, as can the presence of different distracter choices for multiple choice tests. Here, we investigated whether two of the tutors were able to predict which of the items would be most difficult for the quiz takers. This is an attempt to tap into the knowledge that tutors have gained through interaction with students in the teaching process. Knowing the difficulty of the quiz items is important for building a theory of teaching and learning the concepts and for writing assessments at the appropriate level of demand. It is important to highlight discrepancies between what trainers think is difficult and what learners actually find difficult to inform revisions of materials and development of future materials, including those used for face-to-face training.

Research evidence on educational assessment judgments more broadly indicates that neither teachers nor examiners are good judges of the difficulty of items (e.g. Impara and Plake, 1998; Good and Cresswell, 1988; Meyer, 2003). Typical findings on this topic include the results from Wolf *et al.*'s (1995) study, in which a panel of experts were asked to rate the difficulty of 30 mathematics questions. These experts' ratings were then correlated with empirical values of difficulty of the items calculated on the basis of 301 students' performances on the test. The overall correlation was statistically significant, but moderate (approximately 0.4: Table 1, p. 347), indicating that these experts were not able to judge well the order of difficulty of the items for students. Good and Cresswell's (1988) research was more encouraging because their examiners were at least able to judge which question papers were more, and which were less, difficult. The magnitude of the differences in difficulty was not judged well in Good and Cresswell's studies, however.

Figure 12 shows data from Meyer's (2003) study of A-level Economics questions. In her study, A-level examiners were asked to judge the likely success rates on each multiple-choice question for students who would be awarded a grade E on the examination. The thick line on Figure 12 shows the empirical outcomes for students on the examination and each of the other lines represent estimates of the success rates from five senior examiners. Correlations were moderate and, as can be seen, there were discrepancies in the empirical and estimated success rates of students for these questions. Pollitt's work helps elucidate why it is difficult for examiners to judge the demand of questions and distinguishes between the concepts of demand and difficulty.

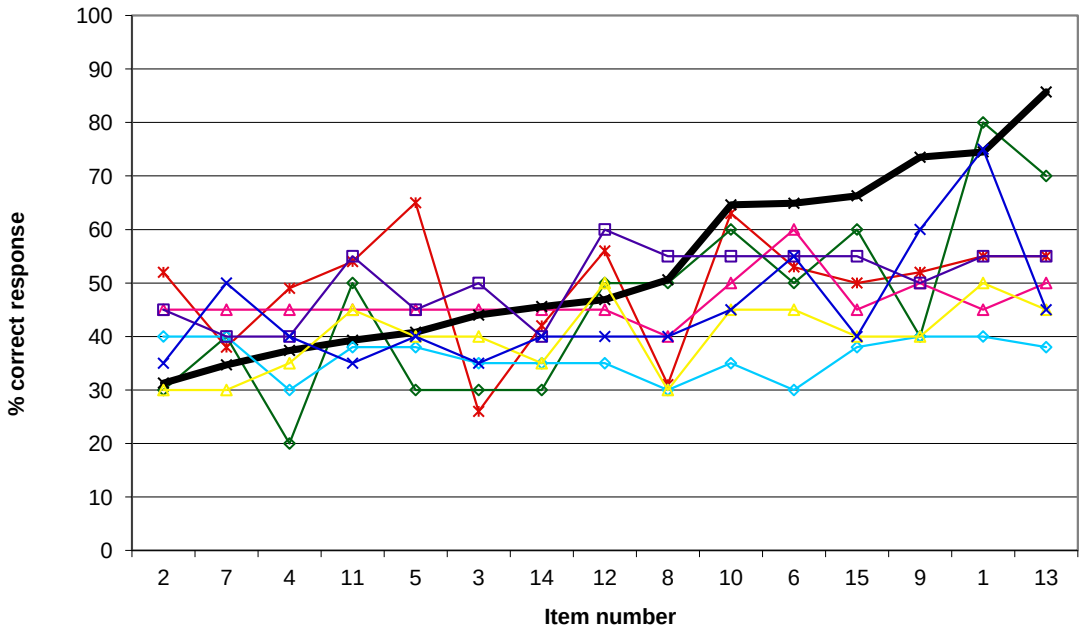


Figure 12 Estimates of success rates on items and the actual outcomes for A-level Economics questions (from Meyer, 2003)

Despite the foregoing research and the widespread nature of examining, knowledge in this area is not at a stage where accurate predictions can be made about the likely difficulty of questions. Nonetheless, tentative recommendations have been made in the literature about which factors should be taken into account (Pollitt *et al.*, 2007b).

Instructions to raters

1. Read the module up to the first quiz
2. Read the items in the quiz
3. Go back to the beginning of the quiz and, for each item, estimate the proportion of learners who would get the item correct. Complete these estimates on the response sheets provided.
4. Continue reading the module up to the next quiz and repeat 2 and 3 above.

Two tutors provided an estimate of the proportion (percentage) of learners who would get each quiz item correct. To standardise the process, instructions were given to have in mind a group of students who were encountering the concepts for the first time. In reality, a diverse group of students (in terms of expertise) attempted the items. Instructions to raters also asked them to try to ignore the effect of item format, and to have in mind a population of students who had worked their way through the materials in sequence and had understood the pre-requisite material for each lesson. Ratings were made for the pre-requisite quiz and the quizzes associated with modules 1 to 5 (149 items in total).

Statistically significant correlations were found between ratings and actual success rates (**Table 5**), but they were moderate for both raters. Interestingly, as in the Meyer (2003) study, the most senior rater had a higher correlation between ratings and success rates (0.62). However, in the Meyer (2003) study, this high correlation was not replicated in the following examination series. Given the number of influences upon success rates, it is possible that experience does produce better estimates of likely success rates, but that it is masked at times by interactions with other features of the assessments. Whilst significant correlations are encouraging, moderate correlations indicate that there is much that subject matter experts are not able to gauge in terms of difficulty of items.

Table 5 Correlation between ratings and statistics on difficulty

	Rater 1	Rater 2
Success rates	0.49	0.62
Rater 2	0.44	

There was a moderate, but significant association between the two raters' estimates of item difficulty (0.44). Raters either perceived factors associated with the items differently, or weighted them differently in making their ratings.

As discussed previously, question format influences the rate at which learners will be successful on an item. The purple dots on **Figure 13** and **Figure 14** represent the numerical response items and are largely distributed below the line, indicating that the raters estimated a higher proportion of success than the actual success rate. This was more so for Rater 1 than Rater 2. Rater 1 under-estimated the success rate on matching response questions: indicated by the green dots lying above the line on **Figure 14**. Estimates for success rates on true or false questions were higher and more accurate from Rater 2.

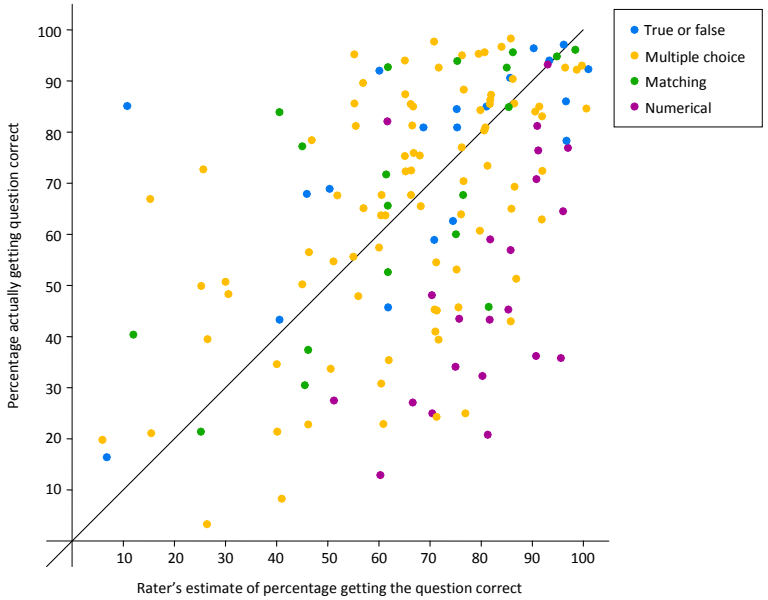


Figure 13 Rater 1's estimated success rates and actual success rates on items

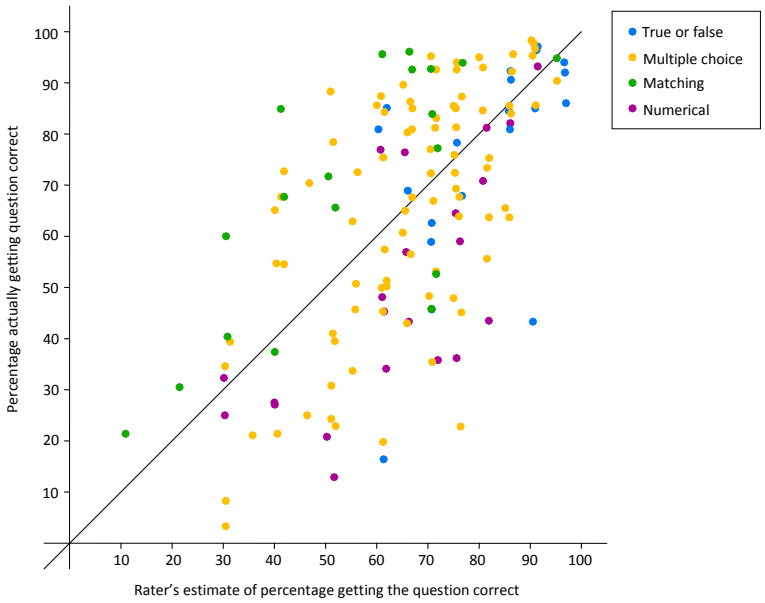


Figure 14 Rater 2's estimated success rates and actual success rates on items

References

Ahmed, A. (2011) *Qualitative Evaluation of LEMMA Questions.* NCRM Report. http://eprints.ncrm.ac.uk/1900/

Ahmed, A. & Pollitt, A. (2010) The Support Model for Interactive Assessment. *Assessment in Education: principles, policy and practice* **17**(2): 133-167.

delMas, R.C. (2004) Mathematical and Statistical Reasoning. Chapter 4 in Ben-Zvi, D. & Garfield, J. (Editors) *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*. Kluwer Academic Publishers, Dortrecht, The Netherlands.

Erduran, S, Simon, S & Osborne, J. (2004) Tapping into argumentation: Developments in the application of Toulmin's Argument Pattern for studying science discourse', *Science Education*, **88**(6): 915-933. ISSN: 0036-8326.

Good, F.J., & Cresswell, M.J. (1988). Grade awarding judgments in differentiated examinations. *British Educational Research Journal*, **14**: 263–281.

Impara, J. C., & Plake, B. S. (1998) Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, **35**: 69–81.

Meyer, L. (2003) Repeat of AQA Standards Unit analyses originally run for the GCE Economics awarding meeting, June 2003. Internal unpublished AQA Standards Unit Paper.

Pollitt, A., Hutchinson, C., Entwhistle, N. and de Luca, C. (1985) *What makes exam questions difficult? An analysis of 'Grade' Questions and Answers*. Edinburgh: Scottish University Press.

Pollitt, A., Ahmed, A. and Crisp, V. (2007) The demands of examination syllabuses and question papers. In Newton, P., Baird, J., Goldstein, H. Patrick, H. and Tymms, P. (Eds) *Techniques for Monitoring the Comparability of Examination Standards.* London: Qualifications and Curriculum Authority.

Rea-Dickins, P, Afitska, O, Yu, G, Erduran, S, Ingram, NR & Olivero, F. Investigating the language factor in school examinations: exploratory studies, SPINE working papers no.2, Report for Study 5.1, for ESRC, DFID, 2009. ISBN: 9781906675912.

Wolf, Lisa F., Smith, Jeffrey K. and Birnbaum, Marilyn E. (1995) Consequence of performance, test, motivation, and mentally taxing items, *Applied Measurement in Education*, **8**(4): 341-351.