Use of mathematical and statistical methods to enhance quality assurance processes for assessment in a pre-clinical veterinary undergraduate course

Brian Catchpole BVetMed PhD MRCVS
Candidate number: 9550151/2

Supervisor: Mr Kim Whittlestone

Word count for body of text (excluding Tables, Figure legends and References)
= 4,483

Research Project Report submitted in part fulfilment of the requirements for the degree of Master of Science in Veterinary Education, The Royal Veterinary College, University of London, 2015.

Declaration

This thesis represents my own work, except where otherwise acknowledged

Brian Catchpole
4th August 2015

Guidelines for Authors

Articles

Articles are the primary presentation mode of communication in the Journal, and are usually between 2500–5000 words in length. All articles must include abstracts, practice points and notes on contributors. Glossary terms should be added if appropriate (see below for further details).

Manuscript Preparation

Manuscripts submitted to Medical Teacher should be written in English and conform to the style guidelines set forth by APA, as per the most recent Publication Manual of the American Psychological Association (6th edition).

Manuscripts should be typed using double-spacing (except tables which should be single-spaced), with margins of at least 2.5 cm (1 inch). All pages should be numbered.

Title page The first page of the manuscript should contain the following information:

i) the title of the paper
ii) a short title not exceeding 45 characters for use as a running head
iii) names of authors
iv) names of the institutions at which the research was conducted
v) name, address, telephone and fax number, and email address of corresponding author.

Abstract All papers should be accompanied by an abstract of up to 200 words. The abstract should reflect the content of the paper including methods used, results, and conclusions drawn.

Text This should in general, but not necessarily, be divided into sections with the headings: ‘Introduction’, ‘Methods’, ‘Results’, ‘Discussion’ and ‘Conclusion’.

Practice Points Up to 5 short bullet points which summarise the key messages of the article should be included (not required for short communications). ‘Practice Points’ will be included in a box at the end of the article.

Notes on Contributors All articles should be accompanied by ‘Notes on contributors’, short biographical notes on each contributor to a maximum of 50 words per contributor.

Glossary Terms If you feel that there are terms or concepts central to your paper that the reader may not be familiar with, please include definition of these terms, giving if possible a reference. Your definitions will then be added in a box at the end of your paper and added to the MedEdWorld glossary.

References References should be in APA 6th format, with names and dates in brackets in the text, and the full reference listed at the end of the paper, in alphabetical order by first author, as follows:

Canadian Institute for Health Information (CIHI) 2009. Supply, distribution and migration of Canadian physicians, 2008. Retrieved July 8, 2010. Available from: http://secure.cihi.ca/cihiweb/products/SMDB_2008_e.pdf

Parmelee DX, Michaelsen LK. 2010. Twelve tips for doing effective team based learning (TBL). Med Teach 32:118–122.

Schmidt HG, Moust JHC. 2000. Factors affecting small-group tutorial learning: A review of research. In: Problem-based learning: A research perspective on learning interactions. Evensen DH, Hmelo CE (eds.). Routledge, Taylor & Francis, Inc.: Kentucky, USA. pp. 19–52.

Illustrations and tables Illustrations and tables should not be inserted in the appropriate place in the text but should be included at the end of the paper, each on a separate page.

Tables should be given Arabic numbers (e.g. Table 3), and their desired position in the text should be indicated. Tables should be used only when they can present information more efficiently than running text. Care should be taken to avoid any arrangement that unduly increases the depth of a table, and the column heads should be made as brief as possible, using abbreviations liberally. Lines of data should not be numbered nor run numbers given unless those numbers are needed for reference in the text. Columns should not contain only one or two entries, nor should the same entry be repeated numerous times consecutively. Units should appear in parentheses in the column heading but not in the body of the table. Words or numerals should be repeated on successive lines; 'ditto' or 'do' should not be used. Tables should be typed using single-spacing.

All photographs, graphs and diagrams should be referred to as Figures and should be numbered consecutively in the text in Arabic numerals (e.g. Figure 3). A list of captions for the figures should be submitted on a separate sheet (or where figures are uploaded as separate files, captions can be entered during the electronic submission process) and should make interpretation possible without reference to the text. Captions should include keys to symbols. Avoid the use of colour and tints for purely aesthetic reasons. Figures should be produced as near to the finished size as possible. All files must be 300 dpi or higher. Please note that it is in the author's interest to provide the highest quality figure format possible.

Please do not hesitate to contact the Publisher’s Production Department if you have any queries.

Acknowledgments and Declaration of Interest sections

Acknowledgments and Declaration of interest sections are different, and each has a specific purpose. The Acknowledgments section details special thanks, personal assistance, and dedications. Contributions from individuals who do not qualify for authorship should also be acknowledged here. Declarations of interest, however, refer to statements of financial support and/or statements ofpotential conflict of interest. Within this section also belongs disclosure of scientific writing assistance (use of an agency or agency/ freelance writer), grant support and numbers, and statements of employment, if applicable.

Acknowledgments section Any acknowledgments authors wish to make should be included in a separate headed section at the end of the manuscript preceding any appendices, and before the references section. Please do not incorporate acknowledgments into notes or biographical notes.

Declaration of Interest section All declarations of interest must be outlined under the subheading “Declaration of interest”. If authors have no declarations of interest to report, this must be explicitly stated. The suggested, but not mandatory, wording in such an instance is: The authors report no declarations of interest. When submitting a paper via ScholarOne Manuscripts, the “Declaration of interest” field is compulsory (authors must either state the disclosures or report that there are none). If this section is left empty authors will not be able to progress with the submission.

Please note: for NIH/Wellcome-funded papers, the grant number(s) must be included in the Declaration of Interest statement.

MEDICAL TEACHER

Use of mathematical and statistical methods to enhance quality assurance processes for assessment in a pre-clinical veterinary undergraduate course

BRIAN CATCHPOLE
Royal Veterinary College, University of London, UK

Abstract

Background: Quality assurance processes are an important, but labour-intensive aspect of assessing pre-clinical veterinary students and making judgements on progression.

Aim: To evaluate the validity of an Ebel method of standard setting for multiple-choice question (MCQ) tests and to investigate whether reliability statistics might be useful for evaluating marking of long answer questions (LAQ).

Methods: Data from MCQ tests administered to first and second year veterinary students in 2010-2015 were evaluated. The Ebel method was compared with other standard setting techniques. Statistical methods were applied to evaluate consistency of marking of LAQ. Staff were surveyed for their views on standard setting and sample marking processes.

Results: The Ebel cut score led to variation in failure rates, which lacked alignment with those in the examination as a whole. There was poor agreement between panellists in predicting question and student performance in the examination. Use of a combination of methods (modified Cohen, Hofstee and linear regression) might be more appropriate for determining the cut score, particularly when used against a historical standard. A 'trustworthiness profile' was designed for evaluating marking of LAQ. In an essay paper, this was demonstrated to be a useful adjunct to sample marking.

Conclusions: Statistical methods could potentially reduce, refine or replace more subjective, labour-intensive quality assurance processes in assessment.

Short title: Statistical methods for assessment

Correspondence: B. Catchpole, Department of Pathology & Pathogen Biology, Royal Veterinary College, Hawkshead lane, North Mymms, Hatfield, Herts, AL9 7TA, UK, Tel: 01707 666388; email: bcatchpole@rvc.ac.uk

Practice points

- There are a number of methods available for standard setting of MCQ examinations, but there is no agreed 'gold standard' method.
- It is important to evaluate whether the method selected for standard setting of MCQ tests is being applied appropriately and demonstrates validity.
- A combination of methods of standard setting, based on analysis of populations of differing ability within the cohort, might provide a more defensible pass mark, compared with the Ebel method.
- Subjective analysis of marker reliability by sample marking can sometimes lead to anomalous results.
- Statistical analysis of marking data for long answer questions can be used to generate a 'trustworthiness profile' that can help inform decision making processes for quality assurance of assessment.

Notes on Contributors

Brian Catchpole is Professor of Companion Animal Immunology in the Department of Pathology and Pathogen Biology at the Royal Veterinary College. He is Strand Leader for the Principles of Science strand of the Bachelor of Veterinary Medicine degree and a Departmental Teaching Coordinator. He is interested in various aspects of assessment of undergraduate veterinary students.

Introduction

Assessment of veterinary students in the UK has come under increasing scrutiny from external stakeholders, such as the Quality Assurance Agency (QAA)[1], as well as professional statutory regulatory bodies, such as the Royal College of Veterinary Surgeons (RCVS)[2], European Association of Establishments for Veterinary Education (EAVE)[3] and the American Veterinary Medical Association (AVMA)[4]. A variety of assessment modalities are often employed during the pre-clinical stage of undergraduate veterinary degree courses, which are designed to assess different aspects of student learning and ability.

Quality assurance (QA) processes, such as standard setting of multiple-choice question (MCQ) tests and double/sample marking of long answer questions, can be time consuming and labour intensive. Furthermore, external examiners, who play an important role in QA processes, are required to scrutinise a substantial amount of assessment material and data in a relatively short time. Thus, there is a requirement to ensure that assessments are reliable and robust, but the procedures involved can come under pressure from time and staff resourcing perspectives.

There is no ‘gold standard’ technique for standard setting, but whichever method is selected, it should be fair, defensible, practical and transparent (Cusimano 1996; Norcini 2003; Zieky et al., 2008). Standard setting methods can be categorised as criterion-referenced (absolute), norm-referenced (relative) and compromise (Cizek and Bunch, 2007). In many instances, the purpose of standard setting is to establish a cut score that can be applied to differentiate between two states of performance (i.e. pass/fail) (Cizek, 1993). However, in other

[1] See: http://www.qaa.ac.uk/en/Publications/Documents/understanding-assessment.pdf

[2] See: http://www.rcvs.org.uk/education/approving-veterinary-degrees/

[3] See: http://www.eaeve.org/about-eaeve/mission-and-objectives.html

[4] See: https://www.avma.org/professionaldevelopment/education/accreditation/colleges/pages/default.aspx

situations, an MCQ test is one component of a much larger examination (in our case, typically contributing ~20% to the final mark) and the purpose of standard setting is not to determine pass/fail status, but rather to scale the marks for inclusion in the final assessment dataset.

Our institution has adopted a criterion-referenced approach to standard setting, based on use of the Ebel method (Ebel, 1972; Case and Swansen, 1998), in which a panel of examiners scrutinise questions and categorise them in terms of three levels of difficulty (Easy, Moderate or Difficult) and three levels of relevance (Essential, Important or Desirable). Panellists are also required to indicate the proportion of questions of each category that a minimally proficient student would be expected to answer correctly. Analysis of data from the panel allows a cut score to be calculated.[5]

Providing empirical evidence in support of the cut score is integral to internal validity, and documenting the impact of applying the cut score on failure rates, as well as the relationship to decisions on other assessments, is an important aspect of external validity (De Champlain 2014). Apart from method comparison studies, there is a lack of research literature on the validity of the Ebel method, when applied under different circumstances. Therefore, the first aims of the project were to evaluate the impact of standard setting on MCQ tests and the effectiveness of the Ebel method, when compared with alternative standard setting techniques.

Other components that contribute to assessment of pre-clinical veterinary students include long answer questions (e.g. problem-solving and essays). Historically, double or sample marking has been employed for QA purposes, but this is somewhat subjective in nature, relatively labour intensive and can lead to a degree of uncertainty in terms of what represents acceptable agreement and how to proceed when there is concern that the marking might be unreliable.

[5] See: www.sagepub.com/cizek/ebel

Different statistical methods have been applied for estimating reliability of assessment data (reviewed by Tisi et al. 2013). Inter-rater reliability can be calculated, using methods such as Cohen’s kappa and Fleiss’ kappa for categorical data, or calculating the intraclass correlation coefficient (ICC; Bartlett and Frost 2008) or concordance correlation coefficient (CCC; Lin 1989) for continuous data. Another commonly used approach is to estimate internal consistency, by use of Cronbach’s alpha (Cronbach 1951). However, one problem with use of such methods at our institution is that students are allowed choice in the examination (typically four from six long answer questions). This additional variability in question selection by each student is a major confounding factor, when attempting to apply reliability statistics. Thus, the second aim of the project was to develop a strategy for evaluating marking data from long answer questions that might be informative to internal examiners, external examiners and examination boards. The final aim of the project was to evaluate staff perceptions and attitudes towards current standard setting and sample marking procedures and to solicit feedback on the potential for improvement.

Methods

Student assessment data for the first and second years of the BVetMed degree course (University of London) from 2010 to 2015 were de-identified before being provided for the study. Proformas (*n* = 7 for each of the two years), completed for the Ebel standard setting process in 2015, were made available, with informed consent from panellists for their use in research. Item analysis[6] was obtained after administration of the MCQ tests (45 questions for Year 1; 60 questions for Year 2).

Data were manipulated/graphed in Microsoft Excel 2010 and imported into GraphPad Prism v6 and IBM SPSS v21 for statistical analysis. Testing for normality was undertaken using the D'Agostino-Pearson test. Parametric tests (one-way ANOVA with Tukey's post-hoc testing for multiple groups or Student's *t* test for two groups) were used for normally distributed data, with non-parametric tests (Kruskal-Wallis test with Dunn's post-hoc testing for multiple groups or Mann-Whitney U test for two groups) used for data that was not normally distributed. Fleiss' kappa statistic was used to assess agreement between multiple raters for categorical variables (Real Statistics Resource Pack for Microsoft Excel [7]).

Staff perceptions of standard setting and sample marking were surveyed using a questionnaire, delivered via Survey Monkey[8] (see Appendix), with ethical approval granted (M2014/0038, date of approval 26th May 2015). Ebel panellists were asked to complete the survey in a pilot study and to provide feedback. This led to minor modification of the questionnaire, prior to surveying academic staff across all departments.

[6] See: http://www.speedwellsoftware.com/exams/multichoice/multichoice-paper-based

[7] See: http://www.real-statistics.com/free-download/real-statistics-resource-pack/

[8] See: https://www.surveymonkey.com/

Results

Standard setting of MCQ tests for a pre-clinical veterinary course

Analysis of historical MCQ test data

There had not been a review of the impact/consequences of using the Ebel standard setting technique at our institution. Data from 2010–2014 were analysed, indicating that applying the Ebel cut score led to marked variability in failure rates (Figure 1). There was a high cut score relative to the mean mark for Year 1 assessments in 2010 and 2011, resulting in high failure rates (24.2% and 35.6%, respectively). This was also evident for the Year 2 assessment in 2012. In contrast, the cut score for Year 1 in 2014 resulted in a relatively low failure rate (3%). This degree of variability raised some concern about the reliability of this standard setting method in our hands.

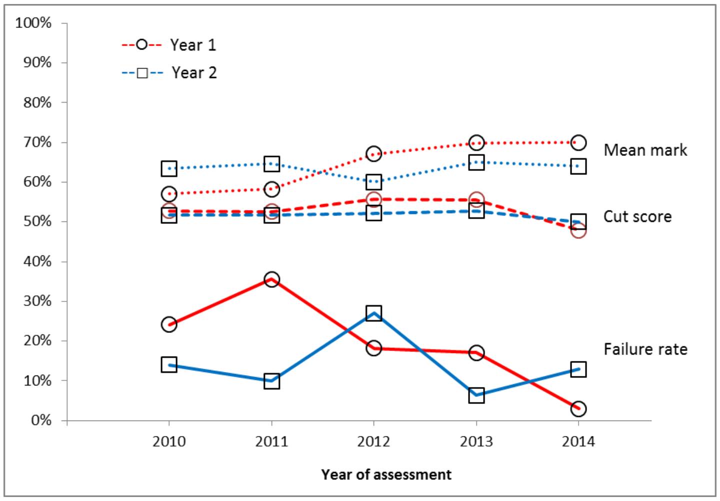


Figure 1. Historical data from MCQ tests, with cut scores generated using the Ebel method of standard setting.

Evaluating the Ebel method

The standard setting process in 2015 was evaluated in terms of internal and external validity (Kane 2001; Hambleton et al. 2012; Cizek 2012, p. 166). Using an approach recommended by Reckase and Chen (2012), each panellist who contributed to the process in determining the final cut score was scrutinised for their individual judgement, based on completion of the Ebel proforma. Considerable variation was observed, in terms of cut scores and consequential failure rates (Figure 2), with Subjects 3 and 7 representing extreme views.

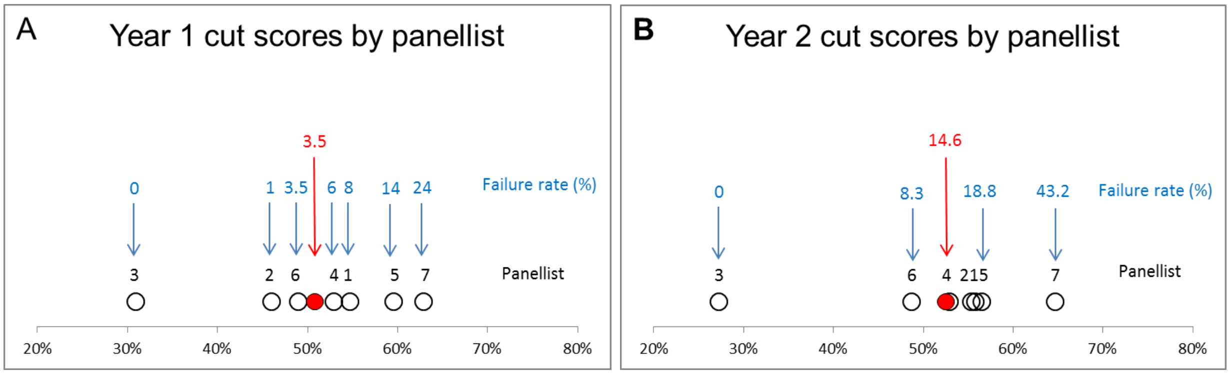
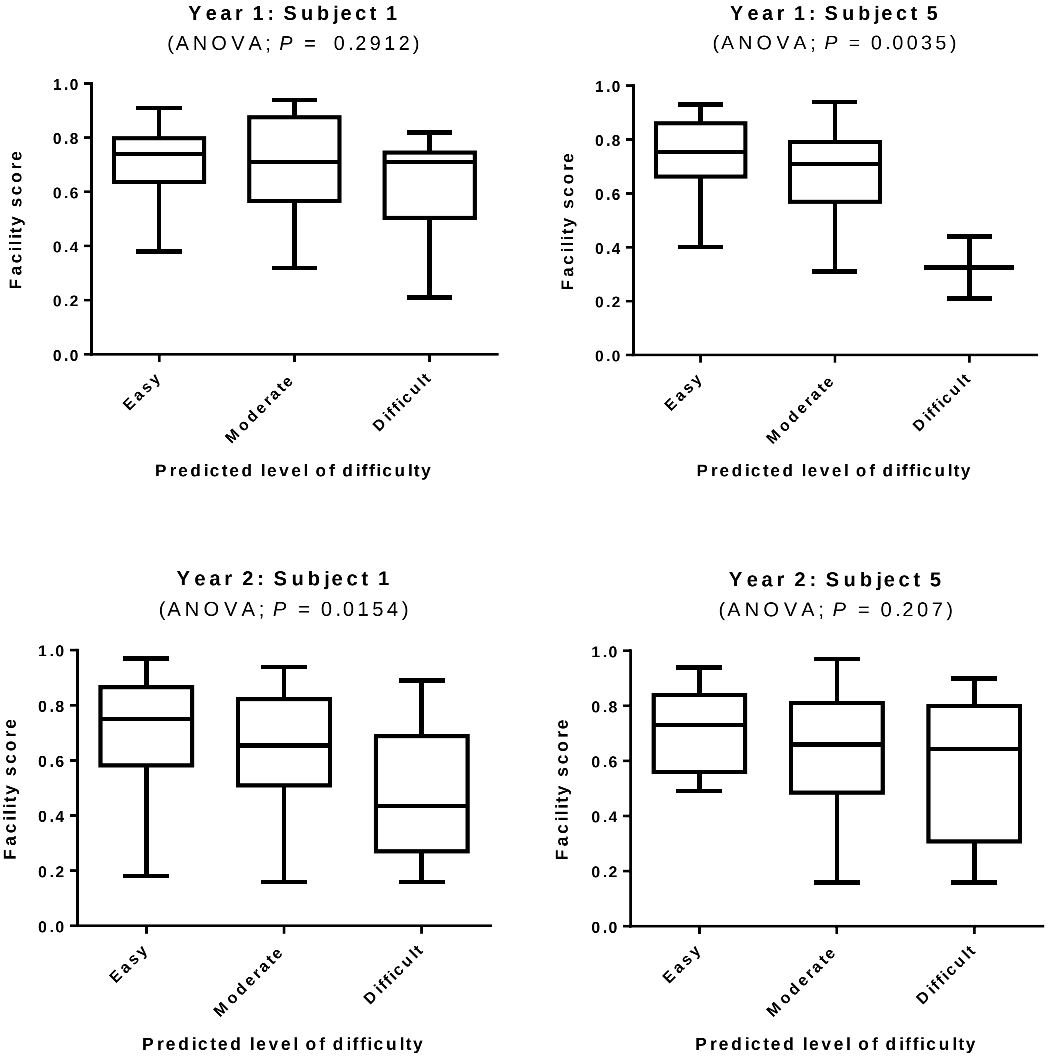


Figure 2. Panellist's contribution to overall cut score for (A) Year 1 and (B) Year 2, based on the Ebel method of standard setting. Red circle represents final cut score. Each individual panellist is identified by a number.

To assess internal validity, inter-rater agreement of question relevance (Essential, Important or Desirable) was assessed and found to be poor (Fleiss' kappa; Year 1 = 7.7%, Year 2 = 8.9%), which did not improve substantially when individuals were systematically removed from the analysis (range; Year 1 = 4.7–10.7%, Year 2 = 5.9–13.3%). Similarly, agreement in categorising question difficulty (Easy, Moderate or Difficult) was also poor (Fleiss' kappa; Year 1 = 7.1%, range 3.95–8.5%; Year 2 = 9.98%, range 7.58–12.2%).

The ability of each panellist to predict question difficulty was evaluated against the facility scores from the MCQ item analysis (the higher the facility score, the easier the question, ranging from 0 to 1). Only one of the seven panellists for each test (Subject 5 for Year 1 and Subject 1 for Year 2; Figure 3) demonstrated significant differences in facility scores for questions categorised according to perceived level of difficulty, indicating that participants were not adept in predicting student performance when answering these MCQs.


Year 1: Subject 1
(ANOVA; *P* = 0.2912)

Year 1: Subject 5
(ANOVA; *P* = 0.0035)

Year 2: Subject 1
(ANOVA; *P* = 0.0154)

Year 2: Subject 5
(ANOVA; *P* = 0.207)

Figure 3. Representative box and whisker plots of facility scores for each category of difficulty predicted during the Ebel standard setting procedure (Subjects 1 and 5 shown). For Year 1, *n* = 45 question; for Year 2, *n* = 60 questions.

Questions were categorised in terms of their difficulty, based on the item analysis and panellists were scored for accuracy (+1 for correct category, -1 for incorrect but adjacent category, -2 for incorrect classification Easy *versus* Difficult). Their scores (Table 1) indicated poor performance in most cases.

Table 1. Accuracy of panellists to predict question difficulty

Panellist	Year 1	Year 2
	(*n* = 45 questions)	(*n* = 60 questions)
1	-18	-3
2	-20	-24
3	-15	-23
4	-12	-31
5	-8	-18
6	-17	-31
7	-25	-22
RANDOM	-25.5	-33

The Ebel method requires panellists to indicate how many questions of each category a borderline passing student would answer correctly. A sample of ~10% of students (*n* = 20) were selected, representing those with borderline passing scores (51.1–57.8% for Year 1, 52.5–56.7% for Year 2; cut scores determined by Ebel method) and their performance evaluated against that predicted for each individual question by each panellist. For first year students, only Subject 5 demonstrated a significant correlation between predicted and actual scores, for second year students, three of the seven panellists (Subjects 1, 3 and 5) demonstrated a significant correlation (Table 2).

Table 2. Ability of panellists to predict borderline student performance

Panellist	Year 1 (*n* = 45 questions)		Year 2 (*n* = 60 questions)	
	r value	P value	r value	P value
1	0.06392	0.6766	**0.4457**	**0.0004**
2	-0.07751	0.6128	0.0313	0.8125
3	0.1786	0.2403	**0.292**	**0.0237**
4	0.2204	0.1458	0.0527	0.689
5	**0.4108**	**0.0051**	**0.2598**	**0.045**
6	0.1931	0.2037	0.21	0.107
7	0.02077	0.8923	-0.05933	0.6525

Evaluation of other standard setting methods

Three alternative methods of standard setting were evaluated, each focusing on a different sub-group within the student population. A modified Cohen method (Taylor, 2011) was applied, based on the formula: Cut score = 0.65 × P_{90} (where P_{90} is the score of the 90th percentile student). A Hofstee method was applied (Hofstee 1983), with the limits set at 45–55% for the acceptable range of the pass mark and 0–20% for the failure rate. On the performance graph, it was noted that there was a high degree of linearity between the 25th and 75th percentiles. Therefore, a new method was designed, based on calculating a cut score from the regression line, when the proportion of students with that score (x axis) is expected to be zero. Figure 4 illustrates how these methods were applied to a representative MCQ dataset.

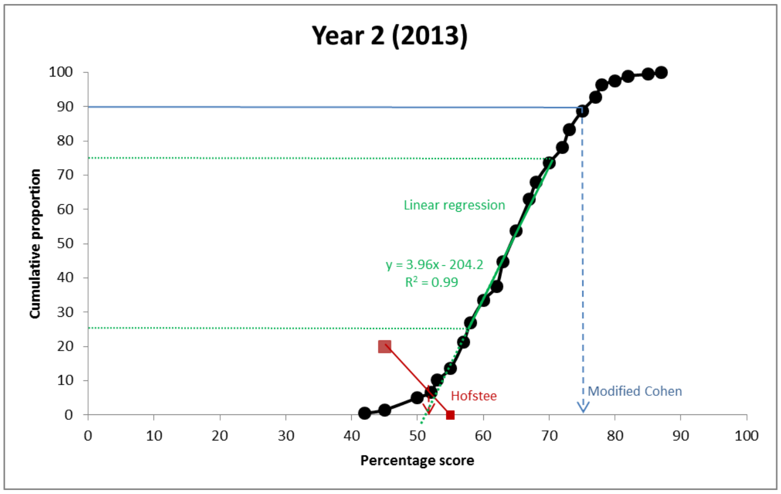


Figure 4. Analysis of a representative dataset of MCQ results using different standard setting techniques, based on the best (modified Cohen), interquartile range (linear regression) and lower (Hofstee) performing students.

These standard setting techniques were applied to the historical dataset to determine what effect they had on cut scores and failure rates (Tables 3 and 4). Taking the mean of the three methods as the cut score (mixed model method; MMM) resulted in more consistent failure rates over the 5 year period, that aligned with the failure rates in the examination as a whole (Figure 5).

Table 3. Application of different standard setting methods to Year 1 MCQ tests

YEAR 1	**2010**	**2011**	**2012**	**2013**	**2014**					
n	198	174	203	210	195					
mean	57.0%	58.2%	67.1%	69.9%	70.0%					
median	57.8%	57.8%	66.7%	71.1%	71.1%	**Failure rates**				
	Cut scores					**2010**	**2011**	**2012**	**2013**	**2014**
Ebel	**52.8%**	**52.6%**	**55.7%**	**55.6%**	**47.9%**	**24.2%**	**35.6%**	**18.2%**	**17.1%**	**3.0%**
Mod Cohen	43.4%	49.1%	52.0%	56.4%	54.9%	7.8%	28.2%	10.8%	17.1%	11.8%
Hofstee	47.8%	44.5%	50.2%	49.7%	50.9%	12.1%	14.4%	7.9%	10.0%	7.7%
Lin Reg	46.1%	38.8%	50.0%	50.4%	53.0%	8.6%	5.2%	7.9%	10.0%	7.7%
MMM	**45.8%**	**44.1%**	**50.7%**	**52.2%**	**52.9%**	**8.6%**	**9.8%**	**7.9%**	**11.4%**	**7.7%**
Failure rate in examination overall						**9.1%**	**12.8%**	**7.7%**	**11.6%**	**5.5%**

Table 4. Application of different standard setting methods to Year 2 MCQ tests

YEAR 2	**2010**	**2011**	**2012**	**2013**	**2014**					
n	179	192	164	197	202					
mean	63.4%	64.6%	60.0%	65.0%	64.0%					
median	63.3%	66.7%	60.0%	65.0%	63.3%	**Failure rates**				
	Cut scores					**2010**	**2011**	**2012**	**2013**	**2014**
Ebel	**51.7%**	**51.7%**	**52.2%**	**52.8%**	**50.0%**	**14.0%**	**9.9%**	**27.0%**	**6.4%**	**12.9%**
Mod Cohen	49.9%	49.9%	49.9%	49.9%	53.1%	11.2%	7.3%	17.7%	1.5%	20.3%
Hofstee	48.8%	50.1%	47.4%	51.8%	48.3%	7.8%	7.3%	13.4%	6.4%	9.9%
Lin Reg	47.4%	51.0%	41.3%	51.6%	45.8%	7.8%	9.9%	4.9%	5.0%	7.4%
MMM	**48.7%**	**50.3%**	**46.2%**	**51.1%**	**49.1%**	**11.2%**	**9.9%**	**10.4%**	**5.0%**	**12.9%**
Failure rate in examination overall						**7.9%**	**3.1%**	**9.6%**	**3.9%**	**9.8%**

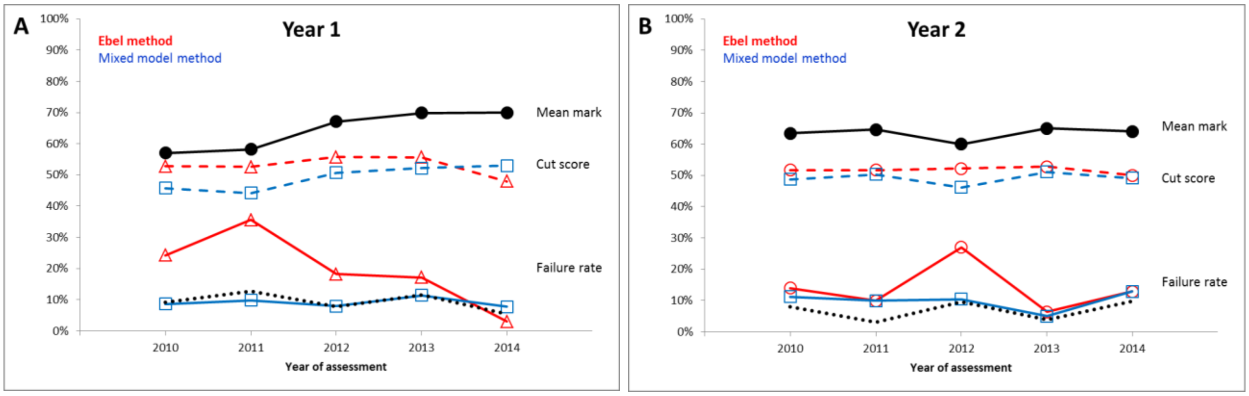


Figure 5. Influence of standard setting methods on failure rates of MCQ tests, comparing the Ebel method with a mixed model method. Black dotted line represents the failure rate for the examination as a whole (22.5% contribution of MCQ to Year 1 examination and 20% contribution of MCQ to Year 2 examination).

Application of the mixed model method of standard setting in pre-clinical veterinary MCQ examinations and use of a 'standard' performance curve

Analysis of historical data demonstrated a significant difference in raw marks between years (Year 1, Mann-Whitney U test *P* <0.0001; Year 2, ANOVA *P* <0.0002), indicating variability in test difficulty and/or ability of different year groups. For Year 2, after scores were normalised for a 50% pass mark, based on use of the MMM, there was no significant difference in scores comparing year groups (*P* = 0.4732), suggesting that this method had corrected for test difficulty, assuming no overall difference in ability from year to year. However, for Year 1, there was a significant difference between year group scores after the standard setting procedure, and post-hoc testing revealed that years 2010 and 2011 were significantly different from 2012–2014, which was likely due to relatively poor performance during these two years (Table 3, Figure 5A). It was decided to combine the MMM-adjusted scores of all students from 2012–2014 for Year 1 (*n* = 608) and from 2010–2014 Year 2 (*n* = 934) to generate 'standard' performance curves for each year group, with a cut score of 50%.

The MCQ results for 2015 were standard set using the Ebel method alongside the MMM (Table 5), showing that applying the MMM cut score led to failure rates that better matched the overall outcome, compared with the Ebel method, which had a very high standard error.

Table 5. Application of standard setting methods to 2015 MCQ tests

2015	**Year 1**	**Year 2**		
n	200	192		
mean	71.0%	65.7%		
median	71.1%	66.7%	**Failure rates**	
	Pass mark		**Yr 1**	**Yr 2**
Ebel	**51.1% (± 4.3%) [a]**	**52.5% (± 4.9%) [a]**	**3.5%**	**14.6%**
Mod Cohen	50.6%	52.0%	3.5%	14.6%
Hofstee	51.7%	49.8%	6.0%	8.3%
Linear regression	55.1%	50.0%	8.0%	8.3%
MMM	**52.5% (± 1.4%) [a]**	**50.6% (± 0.7%) [a]**	**6.0%**	**10.9%**
	Failure rate in examination overall		**7.2%**	**8.7%**

[a] Standard error is shown in parenthesis

For Year 1, there was a significant difference between raw marks and the 3-year standard (*P* = 0.0018). Following normalisation of scores, there remained a significant difference, based on use of the Ebel cut score (51.1%; *P* = 0.0252), but not following normalisation to the MMM cut score (52.5%; *P* = 0.1283; Figure 6A). For Year 2, there were no significant differences between raw scores (*P* = 0.0628; Figure 6B), Ebel-adjusted scores (*P* = 0.9) and MMM-adjusted scores (*P* = 0.1657), compared against the 5-year standard.

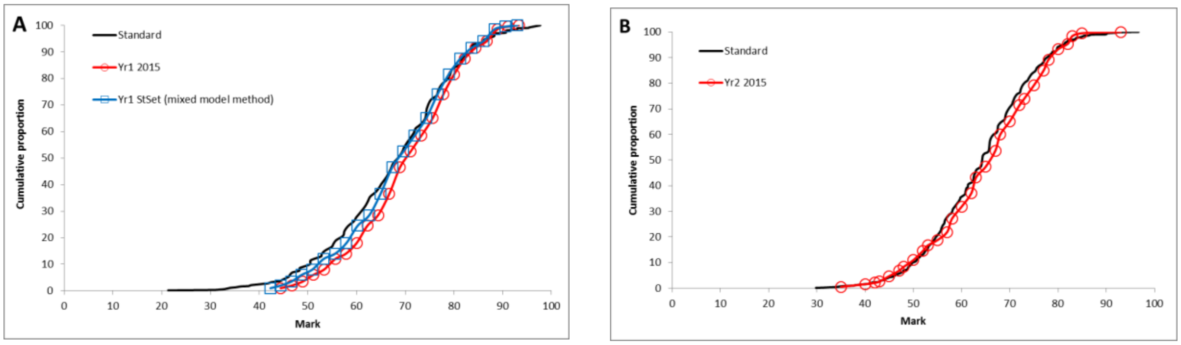


Figure 6. Performance characteristics of MCQ examination in (A) Year 1 and (B) Year 2 against the 'standard' curve. For Year 1 both raw scores and MMM adjusted marks are shown.

Comparing the outcome (pass/fail) based on the final mark and that for the MCQ component in Year 1, there was a high sensitivity (Se) of 98.9% but relatively low specificity (Sp) of 35.7%, after applying the Ebel cut score, compared with the MMM (Se = 97.3%, Sp = 50%), suggesting that the former had less discriminatory capacity in identifying failing students. For Year 2, Se = 91.6%, Sp = 92.9% for the Ebel method, compared with Se = 94.9%, Sp = 85.7% for the MMM, suggesting similar discriminatory capacity.

Improving QA processes for long answer questions on a pre-clinical veterinary course

Long answer questions are routinely sample marked (typically 10% of scripts) at our institution. In the event of a substantial degree of disagreement, scripts are remarked by a third party. The aim of this part of the study was to evaluate whether statistical analysis might be useful in enhancing this process. Although analysis was undertaken on essay and problem-solving question papers for both year groups, only the essay paper of the Year 1 examination will be discussed, as this proved to be problematic in 2015.

Descriptive statistics provided some information (Table 6), and indicated that relatively low scores were awarded for Q6, although this information was not provided to the sample marker, who accepted the marks allocated and did not trigger a re-mark of this question.

Table 6. Descriptive statistics for Year 1 Paper 3 (essays)

	Q1	Q2	Q3	Q4	Q5	Q6
Mean	62.07	63.47	66.20	56.81	58.52	37.07
SD	9.58	10.19	12.63	9.26	10.26	12.02
Median	62.00	65.00	65.00	55.00	62.00	35.00
Upper	82	82	90	82	75	65
Lower	35	35	35	35	27	15
***n* =**	111	155	132	160	54	184

Applying reliability statistics, such as ICC or Cronbach’s alpha, on the study dataset proved to be problematic, since students were allowed choice in the test (two out of three questions from two separate sections). Thus, the variability in student selection of questions and incomplete data for the six questions meant that the equations used to calculate ICC and Cronbach’s alpha were not appropriate. Therefore, a strategy was developed to generate a ‘trustworthiness profile’ for questions in the examination, which could potentially inform decision-making by internal and external examiners.

Data were assessed using the concordance correlation coefficient (CCC; Lin, 1989), with the mark awarded for each question compared with the final mark achieved in the examination. This analysis highlighted that marks for Q4 and Q6 showed poor correlation with the final mark (Table 7). Analysis of the variance of marks for each individual student (i.e. mean of the variance for each student for the test and with individual questions removed) seemed to be informative. The mean variance of student scores in this test was relatively high, compared with previous years, but removal of question 6 had a substantial effect in reducing this (Table 7).

Table 7. Reliability statistics for Year 1 essay paper

Year 1			**Mean variance** [a]
Essay	**CCC** [b]	Overall	182.2
Q1	0.6818	Remove Q1	185.1
Q2	0.4881	Remove Q2	186.9
Q3	0.5658	Remove Q3	182.6
Q4	0.2273	Remove Q4	203.3
Q5	0.4429	Remove Q5	195.2
Q6	0.0690	Remove Q6	66.30
		2010 Overall	121.9
		2011 Overall	110.5
		2012 Overall	89.50
		2013 Overall	104.3
		2014 Overall	85.40

[a] The mean of the variance of scores (*n* = 4 questions) for each student (n = 200) was calculated. This was re-calculated following systematic removal of individual questions from the dataset.
[b] CCC, concordance correlation coefficient (Lin, 1989).

The discriminatory capacity of each question was evaluated, by dividing the year group into quartiles, based on their final marks. Q1 showed good discriminatory capacity, whereas Q4 and Q6 were relatively poor (Figure 7; see Appendix Table S1 for full analysis).

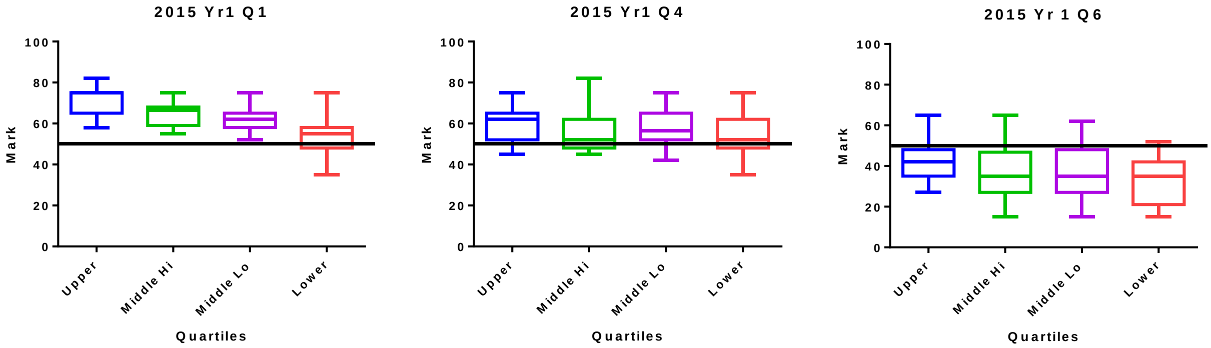


Figure 7. Box and whisker plots illustrating the discriminatory capacity of individual questions, when the year group was divided into quartiles, based on their final mark in the examination. Pass mark of 50% is shown as the horizontal line.

Bland-Altman-style plots were also generated, whereby the difference in marks (Observed – Expected) was plotted against the final mark. Q1 showed the best performance, with a relatively narrow 95% confidence interval centred on zero, although the regression line indicated a trend for over-rewarding poorly performing students, but under-scoring the best performing students (Figure 8). In contrast, Q4 and Q6 both demonstrated profiles suggesting that these questions warrented closer scrutiny (Figure 8).

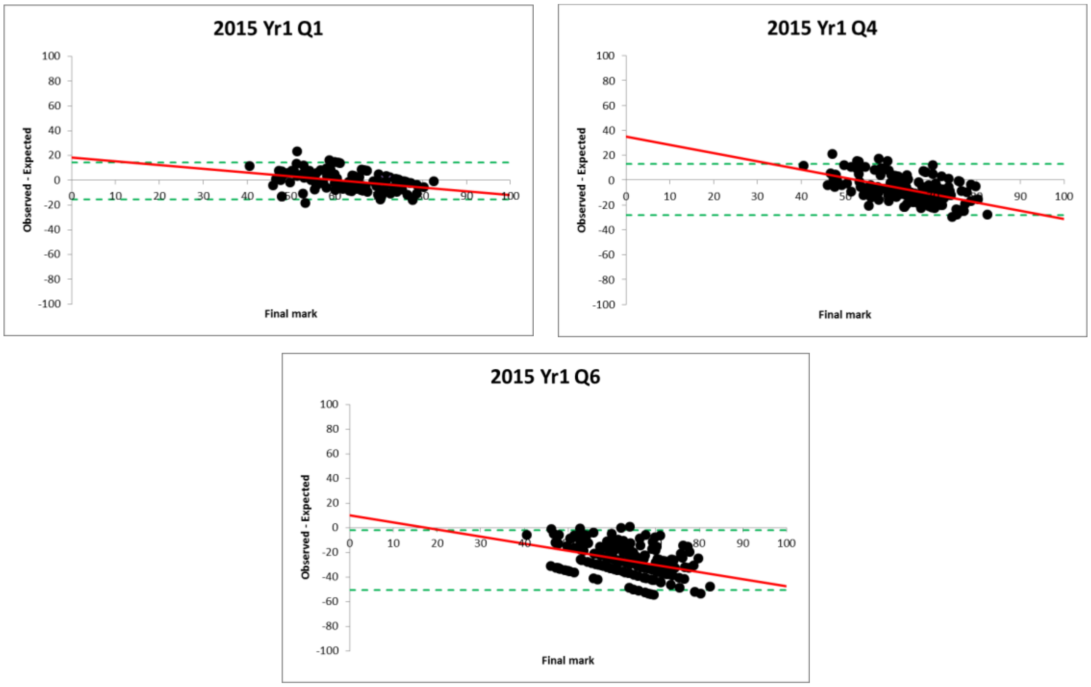


Figure 8. Bland-Altman-style plots illustrating deviation from expected scores (final marks) for Questions 1, 4 and 6. The regression line is shown in red and the 95% confidence intervals are shown by green hatched lines.

Marking of Q5 was considered unreliable by the sample marker and this question was re-marked. However, the statistical analysis indicated that this question demonstrated a reasonable profile and this did not improve substantially by re-marking (Figure 9).

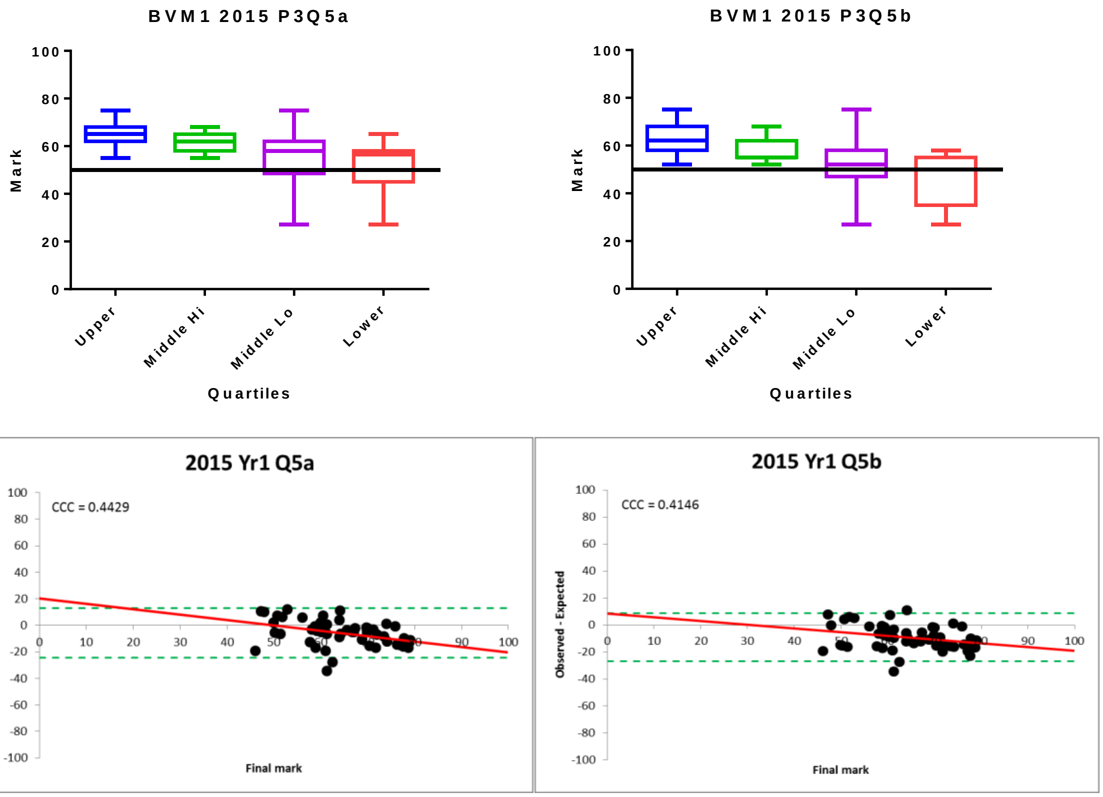


Figure 9. No discernible improvement in the 'trustworthiness profile' for Q5 following re-marking (5a, original marks; 5b, re-mark).

In summary, the analysis indicated that one question in this paper (Q6) required further scrutiny. Although this question satisfied the QA process during sample marking, it was subsequently excluded from the examination at the request of the external examiners. Another question (Q5) was re-marked at the request of the sample marker, although it appears that this marking was satisfactory and re-marking did not substantially improve its performance profile. One further question (Q4) was lacking in discriminatory capacity, showed a relatively low CCC and the Bland-Altman-style plot indicated a negative skew, although this question 'passed' both internal and external scrutiny without comment.

Staff perceptions of standard setting and sample marking procedures

Five of the 7 Ebel panellists completed the questionnaire during the pilot phase of the survey and 50 members of staff completed the final version. Most respondents considered themselves to be moderately (51%) or very (31%) experienced examiners. Ninety three percent agreed that some form of standard setting was required for MCQ tests. However, 56% stated that they felt they had not received adequate training in the Ebel method and that they lacked confidence in assessing question difficulty (80%), relevance (60%) and predicting the performance of borderline passing students (73%). Fifty percent of staff indicated that they lacked confidence that the current standard setting procedure would result in a reliable pass mark. Although there was a lack of awareness of other standard setting techniques (64% of respondents), 73% indicated that alternative methods of standard setting should be considered.

Most respondents (95%) agreed that some form of oversight was required to ensure reliability of marking of long answer questions. Sixty five percent expressed confidence in the current sample marking process and 83% were confident that they were able to evaluate the reliability of marks through sample marking, although 38% were apprehensive in reporting poor reliability. Seventy seven percent indicated that having access to reliability statistics might improve the process and 41% agreed that statistical analysis should be considered as a potential replacement for sample marking.

Discussion

Gathering evidence to support the validity of a standard setting method is an important aspect of their use (Kane 1994). In the current study, there was substantial year-to-year variation in failure rates for the MCQ component of the examination, when Ebel cut scores were applied, which did not align with overall failure rates, suggesting further scrutiny was warranted.

Analysis of data from panellists undertaking Ebel standard setting revealed wide inter-rater variation (Figure 2) and a relatively large standard error in the estimation of cut scores (Table 5). This type of information can be provided as feedback to panellists (Ferrara et al. 2005), stimulating discussion and providing training, as part of an iterative process (Loomis 2012). Reckase and Chen (2012) suggest that those individuals who do not understand the task tend to be at the extremes of the distribution (e.g. Subjects 3 and 7) and this may mean they do not understand one or more key elements of the process. In this respect, Subject 3, who was consistently low in terms of cut scores, indicated performance of the borderline passing student at <20% for several of the categories, yet in a five item MCQ, one would expect a minimum of 20% by guessing.

There was poor agreement between panellists in terms of question categorisation, judging question difficulty and predicting performance of borderline passing students, all of which suggest poor internal validity (Cizek 2012). There are a number of aspects of current practice that should perhaps be reviewed. Fowell et al. (2006) proposed a minimum of 6 participants in standard setting, although Hambleton et al (2012) recommend that “extra panellists should be selected, perhaps twice the number assumed appropriate” (p. 55). In this case, seven members of staff does not seem to be sufficient and probably not optimal. More strategic oversight of the selection process, to ensure broad representation in terms of gender, age and expertise would potentially lead to a more balanced decision-making process (Jaeger

1991). Although training in the Ebel method has been provided, the survey revealed that 56% of staff felt that this was inadequate. This is clearly an area that could be improved, for example through staff development workshops.

Other aspects that could potentially be improved include analysis of performance and feedback to panel members. Reckase and Chen (2012) state that “it seems unethical to have a standard setting process that has one round of ratings with no feedback” (p. 162), which is the current situation. Although there is disagreement in terms of the timing, nature and amount of information provided to panellists, such ‘formative’ feedback (normative and/or consequential) has been shown to improve consistency (Clauser et al. 2002) and to generate a higher level of confidence in the process (Skorupski and Hambleton 2005). A consistent open comment from the survey was that staff had “more confidence with questions in their area of teaching and expertise, but had more difficulty in categorising questions in areas/modules with which they were less familiar”. Provision of item analysis data (particularly showing performance of the lower quintiles) would potentially enable better judgements to be made. Furthermore, providing additional time for discussion, reflection and revision of judgements should be considered.

Since it was not feasible to assess other criterion-referenced methods during the course of the study, selected norm-referenced or compromise methods were evaluated. Combining several methods will not necessarily yield a ‘better’ standard (De Champlain 2014) and there is a wealth of literature illustrating that method comparison studies rarely show agreement (reviewed by Jaeger 1989). However, there does seem to be a rationale for using different methods that focus on performance of subpopulations of differing ability. The MMM offers a ‘cost effective’ solution to standard setting, which might be considered sufficient when the MCQ test forms a component of the examination and the outcome is used to scale marks, rather than to make categorical (pass/fail) decisions. Alternatively, if a criterion-referenced method such as Ebel is considered to be more appropriate, the MMM and standard performance curves could be used as a ‘reality check’ by panellists, chair of the exam board or provided to external examiners as part of the QA process.

The three elements of external validity proposed by Cizek (2012) are that the standard setting procedure should 1) show consistency in terms of repeatability or with other methods of standard setting, 2) show a relationship with decisions made using other sources of information and 3) be reasonable in its outcome. Given the study findings, it is questionable as to whether the current standard setting process is defensible. However, a number of recommendations can be made (see Appendix; Table S1), based on guidelines proposed by Hambleton et al. (2012).

Psychometric methods for evaluating reliability of assessment data are important for ensuring appropriate decision making (Meadows and Billington 2005; Royal and Guskey 2015). Descriptive statistics are often provided during QA processes, but these are limited in their usefulness, when making decisions on marking reliability. In the current study, one question was identified as anomalous, based on low mean/median values, although the relative consistency of the mean and SD of the other questions does not necessarily equate to their reliability (Norman and Eva 2014).

Lucas (1971) demonstrated that for essays, multiple marking significantly increases reliability, but the greatest increase resulted from moving from single to double marking. However, such subjective evaluation of marking (sample, double or team) becomes time consuming and labour intensive with large student groups and it is unclear whether the benefits outweigh the problems (Tisi et al. 2013). Sample marking is more economical, but if this is not blinded, it tends to overestimate marker agreement (Vidal Rodeiro 2007).

Results of the survey revealed a high degree of confidence in current sample marking processes, but there was a degree of apprehension in reporting poor quality marking. This is likely due to social factors impacting on the behaviour of the sample marker in terms of 1) feeling embarrassed at reporting a colleague for poor quality work and 2) feeling uncomfortable about the consequences of a decision that will lead to additional work for others. The majority of respondents indicated that inclusion of reliability statistics into the process would be beneficial.

Reliability statistics are often used for analysis of assessment data (such as Cronbach's alpha), but are problematic when applied to an examination structure whereby students are given choice. Meadows and Billington (2005) found that the problem of low reliability of marking was "exacerbated by the candidates' choice of essay topic" (p. 38). Despite several attempts, it was not possible to find a modification to the Cronbach's alpha equation that would account for the variability in question selection by the examinees. Using an alternative approach, Lin's concordance correlation coefficient seemed to provide useful data and identified two questions (Q4 and Q6) that were worthy of further scrutiny.

Classical test theory is based on the premise that a 'true' score consists of the observed score plus measurement error (Brennan 2011; Tisi et al. 2013). The variance across different test items can be used as an indicator of consistency and likely measurement error. A subtractive approach was undertaken to evaluate how removal of individual questions would impact on the mean variance. Although this seemed a useful exercise, in that Q6 was again identified as an outlier, this approach is somewhat flawed in that removal of one question often left only three data points and therefore a spurious question could impact substantially in evaluation of the others. More sophisticated statistical methods associated with generalizability theory (Cronbach et al 1963; Brennan 2011) or item response theory (Lord 1980; Royal et al. 2014) might be more appropriate for this type of data, but these were beyond the scope of the current study, given the time constraints.

Evaluating the discriminatory capacity of each question is likely to provide information that is useful for decision making. While it is true that some individuals might perform better in certain types of assessment (e.g. MCQ *versus* essays), one would expect that when divided into ability groups, each question should show some discriminatory capacity. Indeed, Norman and Eva (2014) state that “a tool that does not discriminate is useless for assessment” (p. 361). The Bland-Altman-style plot also seems to be a useful addition to the QA process, where clustering of data around the zero line indicates good agreement with the overall assessment of the student’s ability. The negative slope of the regression line seen in most cases likely indicates ‘narrow marking’ of essays by internal examiners, failing to give very low marks for poor responses and inadequately rewarding the best answers. Outputs from this type of analysis could be provided as feedback to internal examiners, helping to inform them of their performance, which is not currently an area of academic activity that is appraised.

Use of reliability statistics could help to inform decision making for QA purposes. Not only would this provide a more objective basis for identifying ‘rogue’ markers (and potentially rewarding reliable examiners) but also for addressing the issue of the rogue sample marker, who might currently accept poor quality marking (a QA issue) or reject acceptable marking (a time and resource issue). According to Meadows and Billington (2005), if essays continue to be viewed as a valuable question format, some degree of unreliability of marking may simply have to be accepted. That said, based on the research findings of this part of the study, a number of recommendations can be made (see Appendix, Table S2).

Conclusions

The study findings raise concern regarding internal and external validity of the Ebel method for standard setting the MCQ tests evaluated in the study. Further investment of resource is required for improving training, feedback and evaluating reliability of cut scores. Considering the limited contribution of the MCQ test to the examination as a whole, more ‘affordable’ normative or compromise methods of standard setting might be considered. Provision of reliability statistics for marking of long answer questions could potentially reduce, refine or replace current QA processes.

Acknowledgements

The author would like to thank the panellists and members of staff who participated in the survey. Particular thanks go to Nicola Toop for her considerable help in accessing the assessment data and to Kim Whittlestone for his advice and helpful comments on the manuscript.

Declaration of interest

The author does not have any financial or personal relationship with other people or organisations that could inappropriately influence or bias the content of the paper.

References

Bartlett JW, Frost C. 2008. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. Ultrasound Obstetrics and Gynecology 31:466–475.

Brennan RL. 2011. Generalizability theory and classical test theory. Applied Measurement in Education 24:1–21.

Case SM, Swanson DB. 1998. Constructing written test questions for the basic and clinical sciences (3rd edition). National Board of Medical Examiners, Philadelphia, PA.

Cizek GJ. 1993. Reconsidering standards and criteria. Journal of Educational Measurement 30:93–106.

Cizek GJ. 2012. The forms and functions of evaluations of the standard setting process. In: Setting Performance Standards: Foundations, Methods and Innovations (2nd edition). GJ Cizek (ed.).: Routledge, Abingdon, UK. pp. 165–178.

Cizek GJ, Buch MB. 2007. Standard setting: A guide to establishing and evaluating performance standards on tests. SAGE Publications, London.

Clauser BE, Swanson DB, Harik P. 2002. Multivariate generalizability analysis of the impact of training and examinee performance information on judgements made in an Angoff-style standard-setting procedure. Journal of Educational Measurement 39:269–290.

Cronbach LJ. 1951. Coefficient alpha and the internal structure of tests. Psychometrika 16:297–334.

Cronbach LJ, Nageswari R, Gleser GC. 1963. Theory of generalizability: A liberation of reliability theory. The British Journal of Statistical Psychology 16:137–163.

Cusimano MD. 1996. Standard setting in medical education. Academic Medicine 71:S112.

De Chaplain AF. 2014. Standard setting methods in medical education. In: Understanding Medical Education: Evidence, Theory and Practice (2nd edition). T Swanwick (ed.).: Wiley Blackwell, Chichester UK. pp. 305–316.

Ebel RL. 1972. Essentials of Educational Measurement. Prentice-Hall, Englewood Cliffs, NJ.

Ferrara S, Johnson E, Chen W. 2005. Vertically articulated performance standards: Logic, procedures and likely classification accuracy. Applied Measurement in Education 18:35–59.

Fowell SL, Fewtrell R, McLaughlin PJ. 2006. Estimating the minimum number of judges required for test-centred standard setting on written assessments: Do discussion and iteration have an influence? Advances in Health Science and Educational Theory and Practice 13:11–24.

Hambleton RK, Pitoniak MJ, Copella JM. 2012. Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results. In: Setting Performance Standards: Foundations, Methods and Innovations (2nd edition). GJ Cizek (ed.).:

Routledge, Abingdon, UK. pp. 47–76.

Hofstee WKB. 1983. The case for compromise in educational selection and grading. In: On Educational Testing. SB Anderson and JS Helmick (eds.).: Jossey-Bass, Washington, DC. pp. 109–127.

Jaeger RM. 1989. Certification of student competence. In: Educational Measurement (3rd edition). RL Linn (ed.).: Macmillan, New York, NY. pp. 485–514.

Jaeger RM. 1991. Selection of judges for standard setting. Educational Measurement, Issues and Practice 10:3–14.

Kane M. 1994. Validating performance standards associated with passing scores. Review of Educational Research 64:425–461.

Kane M. 2001. So much remains the same: Conception and status of validation in setting standards. In: Setting Performance Standards: Concepts, Methods and Perspectives. GJ Cizek (ed.).: Erlbaum, Mahwah, NJ. pp. 53–88.

Lin LI-K. 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45:255–268.

Loomis SC. 2012. Selecting and Training standard setting participants: State of the art policies and procedures. In: Setting Performance Standards: Foundations, Methods and Innovations (2nd edition). GJ Cizek (ed.).: Routledge, Abingdon, UK. pp. 107–134.

Lord FM. 1980. Applications of item response theory to practical testing problems. Mahway, Lawrence Erlbaum Associates Inc., NJ.

Lucas AM. 1971. Multiple marking of a matriculation biology essay question. British Journal of Educational Psychology 41:78–84.

Meadows M, Billington L. 2005. A review of the literature on marking reliability. Retrieved July 24, 2015. Available from: https://orderline.education.gov.uk/gempdf/1849625344/QCDA104983_review_of_the_literature_on_marking_reliability.pdf

Norcini J. 2003. Setting standards on educational tests. Medical Educator 37:464–469.

Norman G, Eva KW. 2014. Quantitative research methods in medical education. In: Understanding Medical Education: Evidence, Theory and Practice (2nd edition,). T Swanwick (ed.).: Wiley Blackwell, Chichester UK. pp. 349–370.

Reckase MD, Chen J. 2012. The role, format and impact of feedback to standard setting panellists. In: Setting Performance Standards: Foundations, Methods and Innovations (2nd edition). GJ Cizek (ed.).: Routledge, Abingdon, UK. pp. 149–164.

Royal KD, Gilliland KO, Kernick ET. 2014. Using Rasch measurement to score, evaluate, and improve examinations in an anatomy course. American Association of Anatomists 7:450–460.

Royal KD, Guskey TR. 2015. Does mathematical precision ensure valid grades? What every veterinary medical educator should know. Journal of Veterinary Medical Education (published ahead of print: DOI: 10.3138/jvme.0115-005R1).

Skorupski WP, Hambleton RK. 2005. What are panellists thinking when they participate in standard setting studies? Applied Measurement in Education 18:233–256.

Taylor CA. 2011. Development of a modified Cohen method of standard setting. Medical Teacher 33:e678–e682.

Tisi J, Whitehouse G, Maughan S, Burdett N. 2013. A review of literature on marking reliability research. Retrieved July 24, 2015. Available from: https://www.nfer.ac.uk/publications/MARK01/MARK01.pdf

Vidal Rodeiro CL. 2007. Agreement between outcomes from different double marking models. Research Matters 4: 28–33. Retrieved July 31, 2015. Available from: http://www.cambridgeassessment.org.uk/ca/digitalAssets/136145_Research_Matters_4_Jun_2007.pdf

Zieky MJ, Perie M, Livingston SA. 2008. Cutscores: A manual for setting standards of performance on educational and occupational tests. Educational Testing Service, Princeton, NJ.

Appendix - Supplementary material

Questionnaire

SurveyMonkey_66395241.pdf

Survey results

Data_All_150803.pdf

Table S1. Discriminatory statistics for Year 1 essay questions

	Q1	Q2	Q3	Q4	Q5	Q6
ANOVA	P <0.0001	P <0.0001	P <0.0001	P <0.0004	P <0.0015	P <0.0007
Upper vs Middle Hi	NS	NS	**	NS	NS	NS
Upper vs Middle Lo	***	****	***	NS	*	NS
Upper vs Lower	****	****	****	***	**	***
Middle Hi vs Middle Lo	NS	**	NS	NS	NS	NS
Middle Hi vs Lower	****	****	****	NS	NS	*
Middle Lo vs Lower	***	NS	***	NS	NS	*

ANOVA with Tukey's post-hoc test applied. NS, not significant.

Table S2. Recommendations for improving Ebel standard setting

- Select a large panel that is representative of the stakeholders. Consider use of pairs of panellists to facilitate discussion after the first round of the process.

- Train panellists to use the method appropriately and assess their performance using historical tests and outcome/impact data to provide feedback.

- Consider providing panellists with normative and consequences feedback during the process. Timing of the panel meeting might need to move after the assessment has taken place if normative data (e.g. Speedwell) is to be provided. Provision of more time for discussion, reflection and revision of judgments before final submission of proformas.

- Conduct an evaluation of the standard setting process and collate responses provided by panellists.

- Compile validity evidence for external examiners, including norm-referenced analysis (MMM) and statistical analysis against historical 'standard'.

Table S3. Recommendations for changes to the QA process of long answer questions

- Analyse assessment data using Lin's concordance correlation coefficient, discriminatory capacity (in quartiles) and prepare Bland-Altman-style plots for each long answer question.

- Meeting of chair of exam board and assessment QA representative to discuss outcomes of data analysis.

- Questions that demonstrate an acceptable standard of 'trustworthiness' are approved and those that require further scrutiny scheduled for sample marking. One approved question per paper also selected for sample marking for quality control purposes.

- Scripts identified for sample marking. 50% randomly selected to represent the range of marks. 50% selected that fall outside the 95% CI on the Bland-Altman plot.

- Sample markers meet to evaluate scripts and feedback to exams office. Current policy employed, dependent upon agreement (approval) or lack of agreement (re-marking required) with first marker.

- Any third marking data analysed to document improvement in profile.

- All reliability statistics and documentation of process submitted to external examiners for scrutiny.