RELIABILITY OF GRADING USING A RUBRIC VERSUS A TRADITIONAL MARKING SCHEME IN STATISTICS

Anthony Morphett[a], Vasileios Giagos[b], Sharon Gunn[a], Jackie Reid[c]

Presenting author: Anthony Morphett (a.morphett@unimelb.edu.au)
[a] School of Mathematics and Statistics, The University of Melbourne, Victoria 3010, Australia
[b] School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University, Manchester M15 6BH, UK
[c] School of Science & Technology, University of New England, Armidale, NSW 2351, Australia

KEYWORDS: assessment, rubrics, reliability, statistics

ABSTRACT

Assessment grading in statistics and mathematics has often been approached in an ad-hoc manner, using marking schemes that attach marks to specific steps of a model solution and often do not explicitly reference assessment criteria. Another approach for grading is to use rubrics. Rubrics are recognised to have several advantages for assessment, but research on the reliability of grading with rubrics is equivocal and mostly conducted in less quantitative disciplines. We present a direct comparison of the reliability of marking of a written statistics assignment using a rubric and using the traditional marking scheme approach. We use a Bayesian statistical analysis and find that both methods yield similar levels of inter-rater and intra-rater reliability.

INTRODUCTION

Grading of assessment in undergraduate statistics and mathematics has often been approached in an ad-hoc manner. A typical approach, which we will call the 'traditional' style, involves lecturers assigning marks to certain steps or components of a task's solution. These marks are often indicated for markers (typically tutors or teaching assistants) with annotations on a set of model solutions. A student's work is graded by deciding whether the student has satisfactorily completed each of the steps to which marks were assigned, and then adding up the total marks for each completed component. Feedback takes the form of a numerical mark, as well as written comments or other annotations on the student work. Assessment in this traditional style is often done without clear identification of desired learning outcomes, and tends to focus on procedural aspects of calculations rather than higher-order skills such as problem-solving and communication (Varsavsky, King, Coady, & Hogeboom, 2014a).

This traditional approach is in contrast to what we will call the rubric-based approach. In this approach, instructors specify in advance the criteria against which an assessment task will be graded. These criteria are stated explicitly for students and markers in a rubric, which also gives descriptors for various levels of achievement for each criterion. An important feature of rubrics is that they state explicitly the criteria against which a piece of work will be assessed, and provide guidelines about what is required for each level of achievement of each criteria, often in the form of descriptors. When assessing student work with a rubric, feedback includes an indication of the student's achievement on each of the criteria, as well as (potentially) written comments linked to the criteria. The levels of achievement in each criterion can be combined to produce an overall grade, for instance by allocating marks or weightings to each of the criteria and their levels of achievement. Rubrics are seen as beneficial for student learning, as they provide transparency in assessment criteria, which can lead to clearer understanding of expectations by students, reduced anxiety, enhanced feedback, and improvements in student self-efficacy and self-regulation (Arter & McTighe, 2001; Reddy & Andrade, 2010; Panadero & Jonsson, 2013). Arter and McTighe (2001) and Brookhart (2013) describe approaches for creating effective rubrics.

There has been increasing focus on assessing with rubrics in undergraduate mathematics and statistics, partly because of their inherent advantages and also because of the growing need for transparency and accountability in assessment practices. One substantial contribution to this discussion was the *mathsassess* project (Varsavsky et al., 2014a), which developed a set of resources for criteria-based assessment in mathematics and statistics. The project also developed 'exemplar' assessment tasks and trialled their use in a range of units at several Australian universities.

A basic consideration when changing assessment practices is how the change will affect the validity and reliability of the assessment. While it is sometimes claimed that the use of rubrics can improve validity and reliability, the evidence for this is equivocal (Jonsson & Svingby, 2007; Reddy & Andrade, 2010). Research on assessing with rubrics has been conducted in various discipline areas including English, management, health sciences and education (Reddy & Andrade, 2010; Panadero & Jonsson, 2013). However, there appears to be little research on rubric use in statistics, particularly for assessing written assignments of the type common in undergraduate statistics, or comparing the rubric-based approach with the traditional marking approach still seen in undergraduate statistics. In this paper, we contribute to this research by presenting a direct comparison of the reliability of marking a written statistics assignment using a rubric and using the traditional marking scheme approach, in terms of absolute agreement (inter-rater reliability) and consistency (intra-rater reliability).

CONTEXT

The research reported here was conducted at a regional Australian university, in an introductory statistics unit for science students. The unit ran from October 2015 to January 2016, and the enrolment was 166 students, mostly first-year science students. The lectures and tutorials were delivered online by a teaching team of one lecturer (the second author) and six tutors. Students were expected to use the statistical software package *R* in the unit. The unit's assessment consisted of 4 written assignments, each worth 10%, as well as online quizzes and a final exam. Standard practice in this unit was that assignments were marked by tutors according to a marking scheme provided by the lecturer. The marking scheme would consist of a sample solution with marks allocated to various steps of the calculations or components of expected answers. An example of such a marking scheme is shown in Figure 3. Tutors would grade student responses by adding up the marks for each successfully completed step or correct answer given by the student.

METHODOLOGY

The third written assignment of the unit (out of 4) was used for this research. Questions for the assignment, and sample solutions, were written by the unit lecturer (the second author). One of the assignment questions (Question 1) is given in Figure 1.

Question 1: Ebola virus fatality rate

In December 2013 the most widespread epidemic of Ebola virus in history begun in Guinea. It quickly spread in several countries of West Africa and is still ongoing. According to a report of the World Health Organization[1] (WHO), as of the end of November 2014 there were 1,327 officially reported deaths out of the 2,164 diagnosed ebola virus cases in Guinea. In epidemiology the term *fatality rate* refers to the proportion of deaths in the reported cases.

(a) Assuming that the reported data are representative of the population of the whole West Africa region, construct a 95% confidence interval for the fatality rate of the Ebola outbreak. Do this by hand and show all your calculations.

(b) Using the confidence interval in (a) and also the fact that the fatality rate of the Malaria disease can reach (the most severe cases) 20%, justify why the Ebola virus poses a more serious threat than Malaria.

(c) The same WHO report announced 3,145 deaths out of 7,635 cases for Liberia, a neighboring country. Using Rcmdr, report a 95% confidence interval for the difference in fatality rates between these two countries. Include both the input and the output of Rcmdr.

Figure 1: Excerpt from assignment questions.

The authors collaborated to construct the marking rubric for the assignment. We drew on the *mathsassess* project (Varsavsky, King, Coady & Hogeboom 2014b) when selecting criteria for the rubric, and the rubric underwent several rounds of feedback and refinement before the final version was completed. The rubric assigned criteria to each sub-question of the assignment, with up to four levels of achievement (Accomplished, Developing, Benchmark, Fail) for each criterion. Each criterion

[1] http://apps.who.int/gho/data/view.ebola-sitrep.ebola-summary-20141202?lang=en

was assigned a weighting, in the form of a number of marks. Two criteria "Clear expression" and "Use of mathematical terminology and notation" were not associated with any specific sub-question, but rather with the assignment overall, and were indicated as such on the rubric. A short description was written for each level of achievement of each criterion, specific to the sub-question in which it appeared; thus, although the criteria were generic, the descriptors gave them a specific interpretation for each sub-question of the assessment task. An excerpt from the rubric, which was associated with Question 1 in Figure 1, is shown in Figure 2.

As the descriptors in the marking rubric often revealed aspects of the required answers or techniques, we could not give the entire rubric to students in advance. Instead, we produced a summary sheet that listed all the criteria used in the rubric, gave generic descriptors of the 'Accomplished' and 'Benchmark' levels of achievement taken from *mathsassess*, and gave the weighting for each criterion based on the marks allocated to it. The summary sheet was supplied to students along with the assignment questions three weeks before the due date. All markers for this subject attended a workshop to train them in the application of the rubric. After the due date, all submissions were marked (for subject assessment purposes) by subject tutors using the complete rubric. The marked assignments, which included the rubric feedback, were returned to the students.

Question	**Criterion**	**Level**				**Mark**
		Accomplished (HD)	Developing	Benchmark (PASS)	Fail	
1(a)	Understanding of key concepts and techniques			1 CI for prop	0 incorrect	/1
	Mathematical manipulations and computations	2 Complete and correct		1 Some minor errors	0	/2
1(b)	Interpretation and explanation of results	1.5 Correctly interpreted a confidence interval as range of plausible values, in context; an evaluation/analysis of what it means, in terms of direction/difference in rates	1 Correctly concludes a difference in rates and provides some explanation; some errors in interpretation of confidence interval	0.5 Conclusion without explanation	0	/1.5
1(c)	Appropriate use of software			1 CI for difference in Rcmdr	0	/1

Figure 2: Excerpt from rubric

In addition to the rubric, a traditional marking scheme was prepared by the lecturer. This marking scheme, based on one used in earlier iterations of the subject, was not used for assessment purposes but was used only for the purposes of this comparative study. An excerpt from the traditional marking scheme for the first assignment question is shown in Figure 3.

To compare the reliability of marking between the rubric-based and traditional marking methods, a sample of 20 assignments was marked using both methods. The sample assignments were randomly selected from the class submissions. In the first stage of this study, five volunteer tutors, all of whom were experienced markers for the subject, marked the same anonymised sample of 20 assignments using the traditional marking scheme. This was done separately to the marking of assignments for assessment purposes, and we ensured that the assignments in the sample had not been marked previously by the markers as part of their normal marking allocation. Two months later, these markers re-marked the same 20 assignments using the rubric. The time delay in the re-mark was to reduce the carry-over effect of markers having previously used the traditional marking scheme for the same sample set of assignments. A limitation of this study is that only one marker had had any prior experience using rubrics, although not in a statistics subject. Future studies involving novice markers and markers experienced with both marking methods would help eliminate experience as a possible confounder.

1 Ebola virus fatality rate **[9]**

(a) We will follow the steps outlined in Section 10.2:

- We calculate the sample estimate of sample proportion $\hat{p} = \frac{1327}{2164} = 0.6132$. [1]
- We calculate the standard error [1]

$$\text{s.e.}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.6132\,(1-0.6132)}{2164}} = \sqrt{\frac{0.2372}{2164}} = 0.0105$$

- The multiplier for a 95% confidence interval is $z^* = 1.96$, or, the approximation $z^* = 2$ is also accepted as correct. [1]
- The 95% confidence interval is $0.6132 \pm 1.960.0105 = (0.5926, 0.6338)$ or $0.6132 \pm 20.0105 = (0.5922, 0.6342)$. [1]

(b) With 95% confidence we can say that the fatality rate of the reported Ebola virus cases in Guinea, up until the end of November of 2014, was in the range of 59.22% to 63.42%. This is a very large fatality rate, three tines bigger that the Malaria's worst rate which makes Ebola virus extremely dangerous. [1]

(c) In Listing 1 we calculate the difference of fatality rates between Guinea and Liberia by issuing the following command in R:

Listing 1 95% Confidence Interval for the difference of fatality rates.

```
> prop.test( c(1327, 3145), c(2164, 7635), conf.level=.95, correct=F)

95 percent confidence interval:
 0.1779969 0.2245981
```

So the C.I is $(0.178, 0.2246)$. [2]

Figure 3: excerpt from the traditional marking scheme

RESULTS

The data set consisted of a mark out of 33 for each of the 20 sample assignments from each of the 5 markers using the traditional marking style, and a mark out of 33 for each sample assignment from each marker using the rubric. A Bayesian statistical analysis was performed on the marks. A series of statistical models was fitted to examine the differences between methods and markers. Initially, a simple random effects model was considered where the effect of the rubric-based approach was compared to the traditional approach. The rubric-based approach introduced a non-significant decrease (0.357 marks, sd 0.2479 and 95% posterior credible interval) to the average mark but this initial comparison did not take into account different characteristics between markers.

To take into account differences between markers, we used a more complex hierarchical model. This model corresponds to an *interaction* model where each method-marker combination is modelled separately. We also took into account differences in students' academic performance; this was modelled with a separate random variable for each student and it was assumed to have the same contribution to the overall mark for all method-marker combinations. We introduced two hyper-parameters, one for each method, to investigate if the assessment methods produce distributions of marks with different variances. The complete model and results are given in Appendix 1. This model allows us to examine the effect that the two *different methods* of marking had on the marks by taking into account differences between *markers* as well as the differences in *academic performance* between students. To investigate inter-rater reliability, systematic differences between markers under the two methods were assessed in terms of their agreement, i.e. if the marks awarded to the same assignment by different markers are close to each other. Intra-rater reliability was assessed in terms of consistency, i.e. if the variability in the marks awarded by a single marker is similar between markers.

The results showed some small but significant (95% credible interval) differences in agreement between markers. The second marker systematically awarded lower marks under both methods, awarding on average 1.08 and 1.92 marks less than the average mark of each script for the traditional and the rubric-based method respectively. The rubric-based method leads to the introduction of a non-zero effect for the third marker. The third marker awarded 1.38 marks less than the average mark of each script under the rubric-based scheme, whereas no significant effect was detected for the third marker under the traditional method. The opposite was the case for the fourth marker; the fourth marker awarded 1.25 marks more than the average mark of each script under the traditional marking scheme but this effect disappeared for this marker under the rubric-based method. We note that these differences between the markers were very small: the marks were on a scale out of 33 (correspond to 3%-5.8%) and have approximately the same magnitude as the variability, which was common between all methods and markers (the corresponding s.d. per student was between 1.275-1.282). No significant effects were detected for the other method-marker combinations.

In terms of intra-marker consistency, the scores of all markers appear to be very consistent for both methods (the estimates for the posterior standard deviations were very similar, a range of 0.524-0.558). This suggests that the variability of the marks does not change from marker to marker nor between the two methods, e.g. the distribution of the marks for one marker may have a different mean compared to the other markers but the variance is similar.

DISCUSSION

The *mathsassess* project (Varsavsky et al., 2014a) conducted several trials of criteria-based assessment, and concluded that 'marking was found to be uniform across all tutors' (p. 24) in most cases. They did not provide quantitative data for this, however, nor compare directly with the traditional marking approach. The analysis presented above suggests that, while rubric-based marking is acceptably reliable between tutors (marker effect sizes ranging from -1.92 to 1.01 marks out of 33, which we consider acceptable for a minor formative assessment task), it is no more reliable than the traditional method (marker effect sizes from -1.1 to 1.3).

Our findings broadly agree with those of Jonsson and Svingby (2007). They report that many studies in their research review obtained levels of reliability that would be considered low by traditional psychometric requirements, but are nevertheless acceptable for the purposes of formative assessment.

Our findings suggest that assessing with rubrics can provide similar levels of reliability to the traditional approach. Thus, concerns about reliability need not be an obstacle to the adoption of rubrics in undergraduate statistics.

CONCLUSION

We found that small but significant differences in the agreement of the markers were present under both marking methods. The introduction of the rubric-based marking had a complex-inconclusive effect on these agreement differences: some were eliminated (marker 4), some were introduced (marker 3), and some were unaffected (marker 2). Both marking methods seem to result in marks with similar variability-consistency. Overall there is no evidence to suggest that there is a difference between the marking methods in terms of reliability.

ACKNOWLEDGEMENTS

Ethics approval (UNE HE15:306) was obtained for this research.

REFERENCES

Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*. Thousand Oaks, California: Corwin Press.

Brookhart, S. (2013). *How to create and use rubrics for formative assessment and grading*. Alexandria, Virginia: ASCD.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review*, *2*(2), 130-144.

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, *9*, 129-144.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch & A. Zeileis (Eds.), *Proceedings of the 3rd international workshop on distributed statistical computing.*

R Core Team (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, URL https://www.R-project.org/

Reddy, Y., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & evaluation in higher education*, *35*(4), 435-448.

Varsavsky, C., King, D., Coady, C., & Hogeboom, K. (2014a). *Developing a Shared Understanding of Assessment Criteria and Standards for Undergraduate Mathematics*. Sydney: Office for Learning and Teaching. Available at https://mathsassess.org/resources/

Varsavsky, C., King, D., Coady, C., & Hogeboom, K. (2014b). *Mathsassess: A Guide to Implementing Criteria Based Assessment in Undergraduate Mathematics.* Sydney: Office for Learning and Teaching. Available at https://mathsassess.org/resources/

APPENDIX 1: STATISTICAL MODEL AND RESULTS

Let y_{ijk} be the recorded mark for the i^{th} method (i = 1: traditional, 2: rubric-based), j^{th} marker ($j = 1, \dots, 5$)and k^{th} student ($k = 1, \dots, 20$). At the first level of the model, we will assume that the marks follow a normal distribution with a mean given from the sum of the following parameters: α_0 ,the overall average mark of both methods, α_{ij}, the average *effect* for the j^{th} marker, α_k, the random effect that expresses the academic performance of each individual student, and ϵ_{ijk}, the error associated with each observation (i^{th} method, j^{th} marker and k^{th} student). We will assume a diffuse prior on the overall average mark by selecting a normal distribution with mean 0 and variance 1000^2 (we note that the marks' range is 0-33). The method-marker effect is assumed to follow a normal distribution with mean 0 and variance ${\sigma_i}^2$, which is different for each method. The observational error and the academic performance effect are assumed to follow a normal distribution with mean 0 and variance ${\sigma_\epsilon}^2$ and ${\sigma_{st}}^2$ respectively. All the variance terms (${\sigma_\epsilon}^2$,${\sigma_i}^2$,${\sigma_{st}}^2$) are assumed to follow a Half-Cauchy (0,1) diffusive prior that allows a very wide range of values. We can summarise the model with the following hierarchical notation:

$$y_{ijk} \sim N(\alpha_0 + \alpha_{ij} + \alpha_k, {\sigma_\epsilon}^2),$$
$$\alpha_0 \sim N(0, 1000^2),\ \alpha_{ij} \sim N(0, {\sigma_i}^2),\ \alpha_k \sim N(0, {\sigma_{st}}^2)$$
$${\sigma_\epsilon}^2, {\sigma_{st}}^2, {\sigma_i}^2 \sim |Cauchy(0,1)|$$

We have used JAGS (Plummer 2003) and R 3.3.1 (R Core Team 2016) to sample from the posterior distribution and the results were summarised in the following Table:

Table 1: Results from Bayesian sampling

Para-meter	Mean	SD	$q_{2.5\%}$	$q_{50\%}$	$q_{97.5\%}$	Para-meter	Mean	SD	$q_{2.5\%}$	$q_{50\%}$	$q_{97.5\%}$
α_0	22.108	1.218	19.723	22.104	24.531	$\alpha_{2,4}$	0.987	0.539	-0.054	0.980	2.070
$\alpha_{1,1}$	0.950	0.539	-0.071	0.934	2.043	$\alpha_{1,5}$	0.476	0.529	-0.544	0.470	1.542
$\alpha_{2,1}$	1.008	0.538	-0.037	1.003	2.079	$\alpha_{2,5}$	0.268	0.537	-0.787	0.268	1.324
$\alpha_{1,2}$	-1.081	0.527	-2.132	-1.074	-0.068	σ_{st}^2	3.304	0.360	2.678	3.278	4.087
$\alpha_{2,2}$	-1.920	0.558	-3.030	-1.913	-0.847	σ_1^2	1.543	1.369	0.352	1.189	4.854
$\alpha_{1,3}$	-0.821	0.524	-1.879	-0.815	0.199	σ_2^2	2.143	1.791	0.566	1.664	6.558
$\alpha_{2,3}$	-1.384	0.553	-2.499	-1.375	-0.316	σ_ϵ^2	26.580	9.263	14.091	24.821	49.496
$\alpha_{1,4}$	1.253	0.543	0.220	1.238	2.361						

In a separate simulation to compare models, we estimated the log-Bayes Factor (-23) of a model that included only the intercept and random effects terms ($\alpha_0 + \alpha_k$) versus the interaction model using the product space method. We biased the prior heavily towards the simpler model ($1 - 10^{10}$) in order to generate a non-trivial posterior sample size. The Bayes factor indicated *strong* evidence towards the more complex (interactions) model.

Proceedings of the Australian Conference on Science and Mathematics Education, Flinders University, 26-28 September 2018, pages 123–128, ISBN Number 978-0-9871834-7-7