Volume 2, Issue 01, January 2024 **ISSN (E):** 2938-379X

THE MOST COMMON PROBLEMS IN ASSESSMENT PLANNING

Karimova Nodira Davronovna
"TIIAMI" National Research University
Teaching Theory and Methodology Department Assisstent Teacher

Abstract:

Learning about resolving common problems means that modes of assessment become less deterministic – real world problems rarely pop up in a form that can be readily solved by an equation or two. This often means the use of projects, portfolios and dissertations. Even simple mathematical problems can give rise to differences of opinion in marking but in the case of the less structured assessments – where answers may be 'better' or 'worse' rather than 'right' or 'wrong' – the assessment issues are much starker. When classes are large and the number of assessors equally substantial there are issues of reliability. This paper looks at some of the issues around the reliability of assessment in such circumstances and uses a large sample, drawn from the marking of Masters dissertations in a related area, to examine issues of double marking. This study suggests a wide disparity between individual markers and that the practice of using a third marker, when the disparity in an individual case falls outside a given range, does not necessarily improve reliability.

Keywords: structured assessment, grade boundary, awarded mark, external supervisors, manipulating the results.

Introduction

Keith Willey and Anne Gardner[1] suggest that [in] an effort to achieve consistent grading between multiple markers, double-blind marking and/or re-marking a random selection of assessment tasks is often undertaken. However, with high student numbers and teaching loads these activities are fast becoming unrealistic. As with Willey and Gradner, many of the issues surrounding marking of dissertations have been concerned with the undergraduate arena where difficulties of marking substantial numbers of project reports are considerable. However, there has been a tendency for postgraduate taught numbers to increase to much the same levels as undergraduate courses. At this level, issues raised include types of assessment peculiar to specific subjects – for example examination of practicum rather than of dissertations. Likewise, some of the more generic studies have focused on the need for detailed marking schemes to increase the accuracy of marking and the validity of various approaches rather than looking at the reliability demonstrated typically by blind double-marking. Raija Kuisma[2] suggests that both intertester and intratester reliability are questionable.' And goes on to suggest that despite very rigorous experiments and experienced markers the reliability is low. However, as with many other researchers, his concern is primarily with the marking of essays

and other assignments throughout a course unit rather than a more substantial piece of work represented by a dissertation or thesis.

In this paper reliability is taken to mean the ability to reproduce the same mark for the same piece of work whereas validity looks at whether the assessment is appropriate to the learning. In terms of reliability, the marks of supervisor and second marker can be seen as estimates of the 'true' mark.

At a November 2010 Examinations Board, some disquiet was expressed by a small number of examiners about the effects that second and third marking were having on their perceptions of the relative merits of their candidates. In this particular university, there is at present a system in place that allows for marks to be averaged when they differ by less than 10 points but for third marking to take place when the difference is greater, or where the resulting average comes close to a grade boundary. The third marker has earlier marks available and this avoids the difficulties that can arise when a third marker marks outside the range of the other two. Although some difference in views on marking can be tolerated, the concerns expressed were that the effect of the second (and sometimes third) marker was to change the rank order of the candidates marked by the supervisor. One supervisor had two candidates awarded the same mark although he had failed one candidate and awarded the other a distinction! This prompted a study of the data from that year's results as well as a wider look at the issues involved.

Sue Bloxham[3] suggests that four assumptions underpin a, largely unchallenged, view of the reliability of assessment "in the higher education community:

1. We can accurately and reliably give a mark to most students' work.
2. Even if individuals' marking may sometimes be inaccurate, internal moderation ensures fair and appropriate standards in marking.
3. Even if internal moderation does not reflect expected standards, external moderation ensures students are assessed against consistent standards across the UK University sector.
4. Students' final award (degree classification) reflects their achievement in a consistent way within and, to a certain extent, across universities."

She then goes on to demolish all of these 'false' premises.

Differences in mark can arise from a number of sources. In this instance the first marker is always the student's supervisor and the mark given may be coloured by the student's performance during the research for the dissertation. Typically this is more likely to be a 'halo' effect where supervisors give higher marks than the written work merits because they have been aware of the effort and thought processes, which they then read into the dissertation although not present.

The Study

The student cohort numbered over three hundred, although some either did not proceed to the dissertation or had the marking suspended pending examination re-sits. A total of ninety-two assessors were employed on the marking process, some marking only two or three dissertations. In the initial analysis it was thought that there were ninety-three assessors but two of them proved to be the same individual with variations in the name used! Many of the

second markers were from a different part of the school with little detailed knowledge of the subject matter of the dissertation.

Results

Tables A1 and A2, in the appendix, summaries some of the results so far. These tables are derived from an analysis between pairs of markers: for each marker the average difference from his or her co-markers has been calculated. A negative difference indicates that this marker generally gives fewer marks than others, whereas a positive score indicates that this individual gives higher marks on average. The range is from one individual who, on average, awarded marks 22% lower than those awarded by others marking the same dissertations, to two individuals who awarded marks on average more than 14% higher. A test was carried out to see if the average difference was in any way correlated with the number of scripts marked. The correlation was found to be very low and, hence, not significant. However, it is noticeable that in the case of the top ten differences the average number of dissertations marked was only 3.

Discussion

There is concern that steps should be taken to reduce the effect of potential marker error. As a result of the initial feedback on this study, the programme team is taking a number of steps to try to reduce the disparity is marks. First, each individual marker will be given advice as to where their marks lie, relative to others. Second, markers will be expected to undertake a larger marking load than three or four dissertations; third, and related to this, the team of markers will be reduced. Finally, permission will be sought to employ some of the 'external' supervisors as part of the second marking team, reducing or eliminating the use of markers from beyond the subject area. The success, or otherwise, of these moves will be monitored and any necessary changes introduced. Initially it had been hoped that a more robust statistical method could be found to reach a final mark, other than using third markers, but the sheer number of markers involved has caused considerable problems of data manipulation, which cannot readily lead to a numerical approach, but which might be resolved in future years. Manipulating the results using the average differences in marks between markers would materially affect the results of about a quarter of the students and there is a judgement to be made as to whether this represents a 'truer' result. Moreover, this approach would not necessarily overcome any individual biases in the supervisor's scoring since 'halo' and 'horns' effects might be averaged out.

Conclusions

The differences in marking, for the same pieces of work, are very substantial – worryingly so. Although the sources of this variation have yet to be determined, the evidence so far is that second markers with a limited knowledge of subject content generally tend to give lower marks. There is no evidence that those with higher marking loads on average mark higher or lower than those with lower marking loads, though those who diverge most from the norms generally have lower marking loads. Although the third marker system can be seen as bringing

some level of consistency, there are still variations within the marking biases of those who undertake this task and, although not conclusive, there is some evidence that third markers detract from the overall inter-assessor reliability.

References

1. K. Willey and A. Gardner, "Improving the standard and consistency of multi-tutor grading in large classes", ATN Assessment Conference, Sydney, Australia, 2010.
2. R. Kuisma, "Criteria Referenced Marking of Written Assignments", Assessment and Evaluation in Higher Education,. Vol. 24, No. 1, pp27-39, 1999.
3. S. Bloxham, "Marking and moderation in the UK: false assumptions and wasted resources", Assessment and Evaluation in Higher Education, Vol. 34, No.2, pp 209-220, 2009.
4. J. Archer and B. McCarthy, "Personal biases in student assessment", Educational Research, Vol. 30, No. 2, pp142-145, 1988.
5. B. McKinstry, H Cameron, R Elton and S Riley, "Leniency and halo effects in marking undergraduate short research projects", BMC Medical Education, Vol. 4, No.28, 2004. Available online at http://www.biomedcentral.com/content/pdf/1472-6920-4-28.pdf
6. C. Edwards, "Assessing What we Value and Valuing What we Assess?" Studies in Continuing Education, Vol. 22, No. 2, pp201-217, 2000.
7. B.Tomkinson and J. Freeman, "Using portfolios for assessment: problems of reliability or standardization", Proceedings of the 30th HERDSA Annual Conference, Adelaide, Australia, 2007.

Web of Teachers: Inderscience Research
webofjournals.com/index.php/1