Objective Analysis of Marker Bias in Higher Education

Subrata Chakraborty[1*], Xujuan Zhou[1], Abdul Hafeez-Baig[1], Raj Gururajan[1], Manoranjan Paul[2], Anuradha Mandal[1], Anila Elizabeth Chacko[1], Prabal D Barua[3]

[1]School of Management & Enterprise, University of Southern Queensland, QLD, Australia
[2]School of Computing and Mathematics, Charles Sturt University, NSW, Australia
[3]Cogninet Australia PTY LTD, NSW, Australia

***Abstract*— Marker bias has been a serious factor contributing to discrepancy in assessments. In this study we analyze one year students' results in a Business Faculty within an Australian university to understand the extent of variation induced by marker bias in multiple marker scenarios. The study shows interesting insights regarding the marking trends of a particular marker, and shows variations among markers in a particular course. The study paves the way for quantification of marker variation through objective analysis.**

***Keywords*— higher education; assessment bias; bias reduction; marker bias; results analysis**

I. INTRODUCTION

Marking bias has been well known source of variations in students' results where multiple markers are used. In order to reduce the marking bias over the years a set of techniques and processes have been developed. To reduce the subjective element in marking, multiple choice questions (MCQ) are often adopted [1]. Although this eliminates the marking bias, MCQ is unable to test students' analytical capacity as it tends to test students' understanding based on memory capacity. Hence in lot of business courses where students' analytical capacity and creativity requires testing MCQ is not a viable solution [2].

Vertical marking approach has been suggested in studies where particular section of the assessment piece is marked by each marker for all the students [3]. This approach works fine if the assessment contains distinct mutually exclusive segments. For example, for an assessment with 5 separate questions to answer, we can use 5 markers each marking one particular question for all the students. This approach improves fairness as individual marker bias is applicable to all students equally. The vertical marking approach however is not suitable if the assessment piece cannot be segmented into mutually exclusive segments and segments cannot be distributed equally among markers. There are some practical challenges with this approach as well. With paper based assessments the markers need to be physically co located otherwise there may not be enough time to rotate the assessments among markers. With time constraint (many education providers have 2 weeks results release policy) even online systems will struggle to coordinate assessment rotation among markers for vertical marking approach.

Self-evaluation and mutual peer evaluation are used in various courses [4][5][6]. However this is often heavily bias and requires heavy moderation by examiner. Often peer assessments are not appropriate in courses where students has the possibility of improving their related task based on ideas gained from peers [7]. This approach would in some cases would be counterproductive for creative work [8].

Formative assessment approach has gained traction with many courses [9][10]. Formative assessments have distinction from summative assessment where continuous improvement process is evaluated [11][12]. Although this is suitable for many courses it may not be ideal for courses run online with limited student teacher personal interactions. Summative assessment is sometimes required to assess student performance based on overall achievements at the end of the course [13].

Marking rubric [14][15][16][17] is a newer approach widely used in higher education sector in Australia and worldwide for summative assessment. The rubric presents a set of guidelines to students and markers outlining the requirements and expectations of the assessment piece. Developing appropriate marking rubric is a challenge itself [18]. If a marking rubric is too prescriptive it reduces misinterpretation but discourages students' creativity and freedom. On the other hand rubric with wider guidelines tend to have different interpretations by different markers. Although a well-developed marking rubric improves consistency between markers bias still exists as different marker may have different tolerance levels to errors made by the students and may apply the guideline in different magnitudes. Marking approaches such as cross marking and feedback discussion are proposed to improve rubrics based marking. In this approach a small number of sample assessments are first marked by all the markers, which are then discussed among the markers and with the examiner so that each marker can adjust their marking process to achieve general consensus.

Moderation process is often adopted to reduce bias in various higher education institutions including Australian higher education sector [19][20]. In moderation process the moderator (examiner or another academic) randomly checks the marked papers and advises adjustments. This process is often done with a small random sample due to high cost and not able to reduce marking bias for all the students.

Although various techniques and approaches have improved marking fairness, academics often feels that marking bias still exists due to marker personality, time of the day marking is done, experience of the marker etc. [17][21][22][23]. In this study we try to provide objective analysis to understand and quantify such bias in an Australian higher education setup. The study highlights how various factors affect variations in marks among students.

II. METHODOLOGY

As shown in Fig. 1, the methodology involves 3 distinct stages including i) Data Collection, ii) Data Cleansing, and iii) Data Analysis.

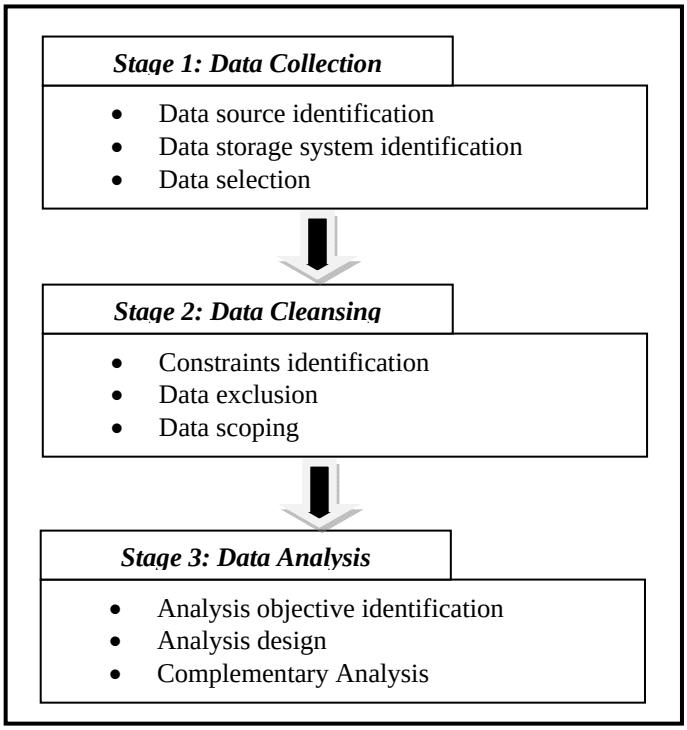

Fig. 1. Different stages of the study methodology

A. Stage 1: Data Collection

In this study we analyze data collected from a business faculty in an Australian university. The university in this study utilizes Moodle learning management system [24] to manage student data, assessment submission, marking and results. From the Moodle system we have extracted student results for year 2015. The data collection was done at the school level at this stage. In the school courses with 2 or more markers were selected for this study from the management and information systems discipline. In these disciplines some courses are offered in multiple semesters. We have considered courses offered in multiple semesters as distinct courses for this study. Based on the data collection scope outlined, we have collected data for 71 courses from undergraduate and post graduate courses. Among the courses 31 courses are from information systems and 40 courses are selected from management courses. Among the information systems courses 9 are from undergraduate level and 22 courses are from postgraduate levels. For the management discipline courses 24 are from undergraduate level and 16 are from post graduate levels. The number of students in different courses vary from 20 to 400. We identified that total of 136 markers were involved in marking of these courses where some markers were involved in marking multiple courses. Although we have collected the data within a limited selected scope, for analysis purpose further cleansing of data was done as described in the following section.

B. Stage 2: Data Cleansing

After initial investigation of the collected data and with some preliminary analysis we identified some key aspects of the data. For comprehensive analysis of the data we excluded some courses from our initially collected data based on several constraints as discussed below:

1) Group assessments: Courses with group assessments were excluded as the group formation were not uniform and marking process is not compatible with individual assessment based marking process.

2) Courses with exam components: Courses where exam is a major component of assessment were not considered in this analysis as the exam markers were not tracked in the results systems individually. Only using low weighted assignment components for courses with exam does not reflect the overall course results hence they were not considered.

3) Markers with limited courses: Markers who marked less than 3 courses were not analysed in depth in this study. Inorder to assess individual marker consistency only markers who marked 3 or more courses were considered in this study.

Applying the data cleansing constraints we have identified 17 markers with more than 3 courses. Among the markers 11 marked 3 courses each, 3 marked 4 courses each, and 3 different markers marked 5, 6 and 7 courses each. The 17 markers were involved in marking 37 different courses. Our analysis is done using these 17 markers and respective 37 courses. To maintain anonymity the 17 markers ar4e coded as Marker 1 to Marker 17. Course names are replaced with generic code as Course 1 to Course 37.

C. Stage 3: Data Analysis

With the selected courses and markers we have conducted three distinct analyses to understand the variations among assessment outcomes. The key objective of these analyses is to identify the bias in marking and perform some quantification of the variation.

1) Course specific data analysis: The marks given by the markers in a particular course is analysed to observe the marking trends of each individual markers. Trend comparison between markers, and trend comparison between individual maker and overall class trend shows the existence of bias. Although the results would be influenced by the student cohort each marker marks, with random student allocation the effect can be assumed evenly distributed. With large student cohorts the trends tend to indicate the general leniency/toughness pattern of any individual marker.

2) Marker specific data analysis: To understand inherent marking consistency of an individual marker, in this analysis several courses marked by same marker is used. Multiple courses and same course in different semesters are considered for analysis where courses are marked by the same marker.

3) Student data analysis: In this analysis the performance trend of a student cohort is compared with the marking trend of a marker for the same cohort. This analysis shows if there is any significant change in student performance trend for the specific marker.

III. Results & Discussions

A. Results for General Statistic

In order to understand the variability in marking we analysed the basic statistics for the selected 37 courses. We calculated Range, Quartiles, Inter quartile range (IQR), Median, Mean, Standard deviation and Coefficient of variation. As summarized in Table 1, we observed wide range of variation among markers in different subjects.

TABLE I. OVERALL VARIABILITY STATISTICS

General statistics on variability	Variation among markers observed over the 37 courses
Range	Up to 70%
Qualtile 1	Up to 15%
Quartile 3	Up to 20%
Median	Up to 20%
Inter quartile range (IQR)	Up to 50%
Mean	Up to 15%
Standard deviation	Up to 60%
Coefficient of variation (Cv)	Up to 50%

The variation in Range has been high but can be attributed to outlier where students did not perform well or students who dropped out. The standard deviation variation indicates there is high variability among markers in some courses. The high variation in IQR indicates that some markers marked in flatter manner than others so that students have received average marks within a narrow band. Although this could be highly dependent on the student cohort, our observation and analysis indicates that some markers indeed have a tendency of narrow band marking highlighting the need for further analysis. The high variation for the Coefficient of variation (Cv) indicates the high variability between some markers in several courses. Our analysis indicates that in all the courses there are variations between markers highlighting the existence of bias among markers. Analysis results in next sections provide a better understanding of the bias among markers.

B. Results for Course Specific Variation

We have completed the trend analysis for 37 courses. Figure 2 shows the trends for Course 11. We can observe that the general trend between markers are similar. However, Marker 8 shows a general tendency of lenient marking compared to Marker 1 and Marker 4. Similar trends were observed in a number of courses highlighting the need to conduct further research how leniency or toughness among markers introduce bias.

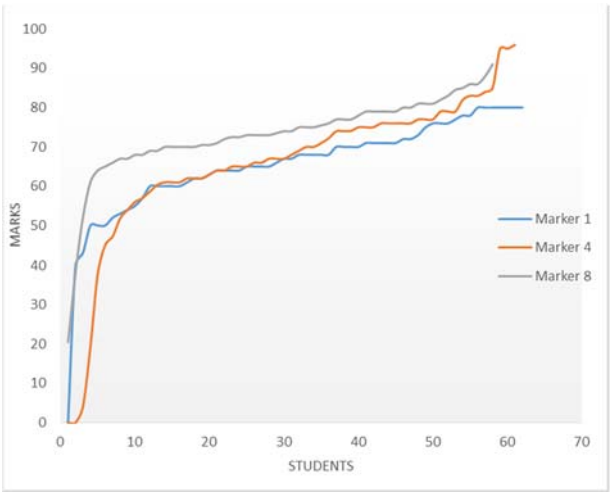

Fig. 2. The trend among 3 markers in Course 11

Several courses showed significantly different trends among markers. Figure 3 shows a sample of such trends for Course 30. Marker 7 and Marker 9 have similar trends which are significantly flatter. On the other hand Marker 3 and Marker 2 shows similar trend which are vastly different from Marker 7 and marker 9. On detail analysis it was identified that Marker 3 is the lecturer for the course and Marker 2 is a very experienced marker. The corresponding box plot shown in Figure 4 shows that Marker 7 and marker 9 has very narrow marking range (IQR) suggesting that due to inexperience these markers may have given marks in the so called safe range between 70-85. This identifies the open issue of how experience affects marker bias. We need further study to find the extent of impact of experience on marker bias.

The course specific analysis shows the variation among markers in a course. Following sections shows results highlighting maker consistency and student performance aspects of bias.

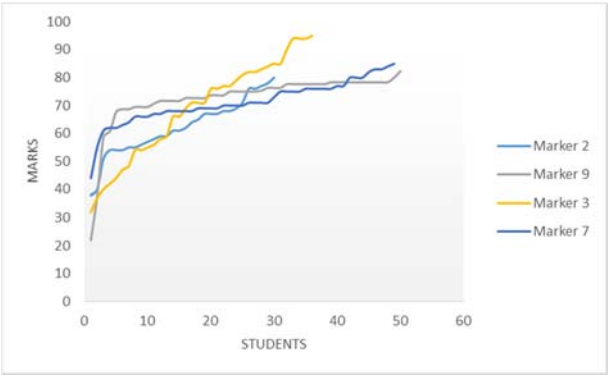

Fig. 3. The trend among 4 markers in Course 30

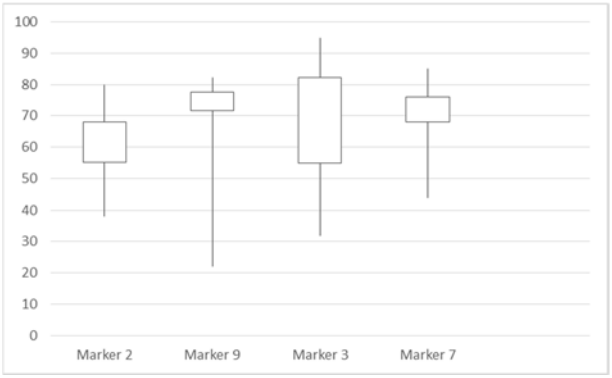

Fig. 4. Box plot for Course 30

C. Results for Marker Specific Variations

We have analysed each of the selected 17 markers who marked more than 3 courses. This analysis is to verify if a marker is consistent inherently. Results show that the markers show general inherent consistency over multiple courses. Although minor variation observed they can be due to several factors such as course content, study level (undergraduate, postgraduate) and student cohort performance. Figure 5 shows the histogram trend for Marker 10. The histogram was based on the grading scale used by the university. Marker 10 marked 5 courses in 2015. The analysis shows that Marker 10 has been consistent over multiple courses, hence inherently consistent in marking. Our analysis with other markers show similar outcomes.

These results suggests that with marker specific variation negligible in this study we need to put more focus on the variations among markers as discussed in earlier section. In the following section we discuss the results for student cohort variation to understand the impact of student performances on marker bias.

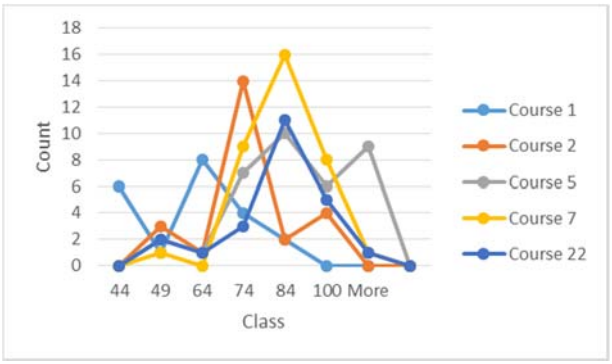

Fig. 5. Marker 10 histogram trend for 5 courses

D. Student Cohort variation

The cohort variation is to understand if a student is affected by being marked by specific marker. The analysis compares student performances in two different assessment pieces within a course. As shown in Figure 6 we have compared each student's marks in two assessments for Course 22. Close comparison shows that students' performances in two distinct assessments were comparable. Although the assessments may be marked by different or same markers (allocated randomly) there is no significant variation between most of the students' performances in two assessments. Figure 7 shows the trends for the students in Course 22 confirming the finding of Figure 6. Similar results were observed for 37 other courses. The results indicate that fluctuations in student performance contributes very less in student marks variation, thus highlighting the impact of marker bias in students' marks variation.

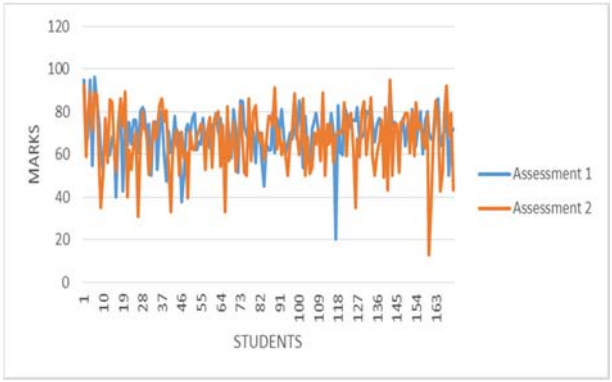

Fig. 6. Student wise comparison between two assessments for Course 22

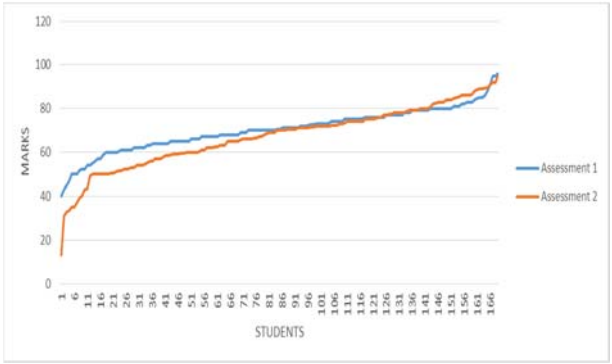

Fig. 7. Trend comparison between results in two assessment of Course 22

IV. Conclusions

In this study we have conducted objective analysis of marks variation. The study highlights that consistency of individual marker and fluctuation of students' performances are weak factor in marks variation among students in a course. Rather the marker bias turns out to be the key factor in marks variation. There is further need for study to scrutinize individual student performance fluctuations using long tern performance such as interim GPA to quantify the impact of student performance on marks variation. The individual marker consistency requires further study using long term trends of individual marker over multiple courses.

The study establishes that there is significant impact of marker bias on marks obtained by students. It also shows that markers have different marking patterns such as lenient or tough, experience plays a part in marker bias. This study

identifies significant future study directions to quantify the bias among markers which can eventually be used to normalize the variation using comparative scale.

REFERENCES

[1] Brown, G.T., Irving, S.E. and Keegan, P.J., 2007. An introduction to educational assessment, measurement and evaluation: Improving the quality of teacher-based assessment. Pearson Education New Zealand.

[2] Scouller, K., 1998. The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. Higher Education, 35(4), pp.453-472.

[3] Zhifang, D., Hehong, F., Meng, Z., Lihua, Y., Jing, S., Qilong, W. and Yongming, T., 2014, December. Individual evaluation for freshman in small size group. In Teaching, Assessment and Learning (TALE), 2014 International Conference on (pp. 453-456). IEEE.

[4] Topping, K., 1998. Peer assessment between students in colleges and universities. Review of educational Research, 68(3), pp.249-276.

[5] Dochy, F.J.R.C., Segers, M. and Sluijsmans, D., 1999. The use of self-, peer and co-assessment in higher education: A review. Studies in Higher education, 24(3), pp.331-350.

[6] Hanrahan, S.J. and Isaacs, G., 2001. Assessing self-and peer-assessment: The students' views. Higher education research and development, 20(1), pp.53-70.

[7] Stefani, L.A., 1994. Peer, self and tutor assessment: relative reliabilities.Studies in Higher Education, 19(1), pp.69-75.

[8] Yorke, M., 2003. Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice. Higher education, 45(4), pp.477-501.

[9] Gibbs, G., 1999. Using Assessment Strategically to Change the Way Students. Assessment matters in higher education, 41.

[10] Bloom, B.S., 1971. Handbook on formative and summative evaluation of student learning.

[11] Harlen, W., 2006. On the relationship between assessment for formative and summative purposes. Assessment and learning, 2, pp.95-110.

[12] Yorke, M., 2011. Summative assessment: dealing with the 'measurement fallacy'. Studies in Higher Education, 36(3), pp.251-273.

[13] Crespo, R.M., Najjar, J., Derntl, M., Leony, D., Neumann, S., Oberhuemer, P., Totschnig, M., Simon, B., Gutierrez, I. and Kloos, C.D., 2010, April. Aligning assessment with learning outcomes in outcome-based education. In IEEE EDUCON 2010 Conference (pp. 1239-1246). IEEE.

[14] Chow, T., Ko, E., Li, C. and Zhou, C., 2012, August. The systematic development of rubrics in assessing engineering learning outcomes. InTeaching, Assessment and Learning for Engineering (TALE), 2012 IEEE International Conference on (pp. T1A-1). IEEE.

[15] Cain, A., 2013, August. Developing assessment criteria for portfolio assessed introductory programming. In Teaching, Assessment and Learning for Engineering (TALE), 2013 IEEE International Conference on (pp. 55-60). IEEE.

[16] McKenzie, S. and Wood-Bradley, G., 2014, December. Using rubrics in IT: Experiences of assessment and feedback at Deakin University. In Teaching, Assessment and Learning (TALE), 2014 International Conference on (pp. 474-479). IEEE.

[17] Buragga, K.A., Khan, A.R. and Zaman, N., 2013, August. Rubric based assessment plan implementation for Computer Science program: A practical approach. In Teaching, Assessment and Learning for Engineering (TALE), 2013 IEEE International Conference on (pp. 551-555). IEEE.

[18] Czaplinski, I., Senadji, B., Adie, L. and Beutel, D., 2014, December. Analysis of moderation practices in a large STEM-focused faculty. In Teaching, Assessment and Learning (TALE), 2014 International Conference on (pp. 346-350). IEEE.

[19] Orr, S., 2007. Assessment moderation: constructing the marks and constructing the students. Assessment & Evaluation in Higher Education,32(6), pp.645-656.

[20] Malouff, J., 2008. Bias in grading. College Teaching, 56(3), pp.191-192.

[21] Greatorex, J. and Bell, J., 2004. Does the gender of examiners influence their marking?. Research in Education, 71(1), pp.25-36.

[22] Merritt, D.J., 2008. Bias, the brain, and student evaluations of teaching. . John's L. Rev., 82, p.235.

[23] Dewberry, C., 2001. Performance disparities between whites and ethnic minorities: Real differences or assessment bias?. Journal of Occupational and Organizational Psychology, 74(5), pp.659-673.

[24] moodle.org. 2016. *Moodle Pty Ltd*. [ONLINE] Available at: https://moodle.org. [Accessed 24 October 2016].