IMPROVING FAIRNESS ON STUDENTS' OVERALL MARKS VIA DYNAMIC RESELECTION OF ASSESSORS

Zhuhan Jiang[1] and Jiansheng Huang[1]

[1]School of Computing, Engineering and Mathematics, Western Sydney University, New South Wales, Australia

ABSTRACT

A fundamental subject delivered at the tertiary level could have a cohort of several hundreds of students distributed into multiple campuses. The running of such a unit typically calls for a teaching team of which a major task is to fairly mark all students' various assessment items. It is well observed that a given assessment is likely to receive different marks if it is given to different markers, often regardless of how detailed the marking criteria are, especially when the content is of subjective or opinion based nature. In this work, we propose an effective strategy to improve the fairness on the students' overall marks by accepting that markers may have inherent marking leniency of different magnitude and by dynamically reselecting markers for different groups of students in such a way that the students will eventually share a similar amount of marking leniency in their overall marks. This strategy is completely objective, purely based on the markers' previous marking statistics, and is independent of the design and interpretation of the marking criteria.

KEYWORDS

Fairer overall marks, marker reselection, objective strategies, effective algorithms.

1. INTRODUCTION

The burgeoning of e-learning or blended learning has brought about radical changes to the education sector in recent years, ranging from the fast delivery of all forms of learning materials and interactive activities, to the more effective learning models and teaching pedagogies [1-3]. An important component in any formal learning is to conduct various assessments on the students and give their marks and marking feedback accordingly [4]. Hence, it's crucial for all students to receive fair and equitable assessment marks on the work they have done [5].

When a student enrols into a course or subject to study theories or technologies, he will have a set of expectations on what knowledge or skills are to be gained and what form of accreditation or certification could be endowed for the learning achievement. Since the topic and skill coverage within the subject in general complies quite well with the students' expectations, how a subject delivery is conducted impacts a great deal more on the student satisfaction on taking the subject or unit, especially for subjects that are offered to massive number of students [6,7] online or concurrently on multiple campuses. More specifically, the equity or fairness of assessments across the students, as well as the individual's expectation on the grade outcome, will be the major factors affecting the students' satisfaction rate or the students' review on the teaching performance [8,9]. The consistency and fairness on the student assessments are also crucial to keeping students' positive learning attitude and motivation. This is in general much less an issue when the class size is small as the total number of involved disparate teaching staff would be small too. For a large cohort of students, different campus lecturers and many different tutors or markers will make the assessment consistency not an easy task at all [10,11,6]. Among many popular assessment pedagogies, there is currently also a trend to adopt peer assessments [12,13] or even automated feedback and marking system [14] somewhere. This may be partly because for large classes peer assessments or the like implicitly reduce the assessment marking tasks by the instructors and tutors. Of course, when peer assessments are heavily involved it may be even harder to maintain the consistency and fairness across the students although it may in principle be

DOI :10.5121/ijite.2017.6201

incorporated into our later proposed marker selection strategy as long as we treat these marking peers just like the formal markers.

The marking inconsistency is typically due to the subjective assessment of certain items, in that what marks should be given is not unequivocally set by the marking criteria. In theory the more detailed the marking criteria are, the less marking inconsistency there will be. However it's simply not possible to specify all potential circumstances in the marking criteria, and therefore subjective assessment is unavoidable. The subjective tendencies, or "biases", are often inherent of a person's personality regardless of how hard a person may try to be fair. Hence it's best to accept that marking biases are unavoidable and then try to find ways to alleviate their impact rather than pretending that such biases can be eradicated completely. To study any tertiary subject, a student typically needs to submit several assessments due on different dates. We propose to explore the marking statistics up to the last marked assessment to estimate how much marking leniency or bias a student may have "suffered" this far, and then utilize the predicted marking profiles to reselect markers for the next assessment so that the predicted accumulated biases will be pro-rata evenly distributed among the students. An initial effort has already been made [15] in this regard where markers are reallocated *student-wise* via algorithms utilizing means and standard deviations for various assessments. This means the reallocation of marking duties will be student-wise rather than group or class-wise, making its administration and in-class marking feedback inconvenient or impossible. One of the main purposes of this work is thus to design a strategy to reassign students group or class-wise to the markers, with a neater algorithm, so as to achieve fairer overall marks for all students. We will also develop an approach to synchronize marking profiles established from different cohorts of students in the past so that proper candidate markers may be better selected for future marking consistency. These extra components of teaching support may be eventually implemented as certain in-house tools similar to those in [16] that have been gradually developed largely on a need-to basis to complement the generic support provided by the web system currently adopted by the University.To recap the main purpose of this work, we reiterate that the running of a large tertiary subject requires in general the support of many assessors to mark or grade the student work, and these assessors will almost inevitably exhibit inconsistent marking leniency no matter how detailed the marking criteria are formulated, especially on the matters of subjective nature. Our proposed solution is therefore to reallocate the assessors objectively for theseparatepieces of submitted student work so that each student will share an estimated equal amount of accumulated marking leniency overall, thus leading to a fairer and more balanced total assessment mark for each student.

This work is organized as follows. We first in section II explore the pitfalls of subjective biases and typical various traditional strategies to reduce the marking discrepancies. We then in Section III set up a proper framework to deal with group-wise marking allocations to assessors of inhomogeneous marking profiles, and propose a simpler but more pertinent and efficient method to reallocate the markers for the different assessments so as to even out the total accumulated marking discrepancies as much as possible, with the corresponding implementation examples demonstrated in Section IV. Section V then extends the use of marking profiles to cross-cohort and multi-cohort profiles so as to improve the initial or later supplementary selection of the markers for better marking consistency and fairness. The conclusions are finally summarized in Section VI.

2. THE PITFALLS OF SUBJECTIVE BIASES

Fairness in assessing student work is essentially embedded into the marking or grading itself, and it has been a topic under consideration for decades if not for centuries. A variety of teaching and assessment methodologies, many in common sense and mostly independent of the use of technologies, have already been in existence. Our consideration of fairness here will lay stress on the equity across the students, rather than whether it is fair to assess students in a particular form

or on a particular topic. The assessment equity [17, 18] is often more subjective than the concept of what and how to assess, where a vast number of contributory factors could range from the type of assessments (written, oral, practical, peer, formative, group), to the topic complexity and pertinence (fundamental, challenging), and further to the extent of marking feedback (evaluative, descriptive). There are also many factors that can affect the assessment equity among the students. Although formal marking criteria, possibly with the use of marking rubrics, can go a long way to make marking or grading accountable, inherent individual "biases" are much harder to defend against. Such biases are often demonstrated by an essentially the same work receiving quite different marks when marked by different markers due to largely the markers' individual or subjective inclinations. This type of biases is generic and can contribute significantly to the breach of the fairness to the students' work and efforts.

With the technological advances in e-education, it has now become possible to reduce those biases more objectively. As the first part of the undertaking, we assume that a teaching instructor would pre-empt the potential partiality as much as possible by going through some of the following strategies:

1. Formalize the assessment criteria as much as possible. This will obviously reduce the mark fluctuation range for any given assessment to mark. In other words, if the same assessment item is given to different markers, the difference of the marks this way is likely to be smaller.
2. Subjective criteria should be replaced by the alternative objective criteria as much as possible. For instance, instead of asking a marker to rate a programming style to see if it's excellent or good or something else, the marking criteria might specify objectively what features are considered good, and therefore how good a program is can be more objectively determined by counting the number of those desirable features. Some form of auto-marking, such as that for the programming tasks [16], can be of good assistance.

It's not difficult to imagine that some markers are just by nature more lenient than the others. Although additional training on the markers can help to a good extent, it's not always feasible or practical. Hence if different assessment items for the same student get to be marked by *suitably* different markers of a given marking team, then it's possible for the students to get a similar amount of marking leniency on the whole and therefore get fairer overall marks. The simplest yet very effective strategy is to swap markers properly to make each student have a fair share of the marking leniency. An algorithm can be formulated [15] to assign a suitable marker for each individual student to mark the next assessment item so that the total accumulated leniency is evened out as much as possible after such a re-assignment of the markers. This approach is however individual-based, and the reallocated marking duties could go across a number of different tutorial groups or even campuses, and therefore create considerable extra management work on both the unit coordinator and the markers. Moreover, students may share some common problems associated with a specific group, and a group-based reallocation would allow more conveniently the relevant tutor for the group to address such problems. For these reasons we will develop below strategies that will make full use of the fact that the markers will always be assigned to mark assessments group or class-wise, where the group sizes are largely fixed due to the fixed class or lab room size. It is perhaps important to note here that swapping markers is much better a choice than, say, rescaling different groups of students differently according to who their markers are. This is because it is often impossible to formally justify the use of different rescaling for the different groups to the students, while in the case of reallocating markers all existing marks can still be deemed "absolutely" correct on the papers.

3. IMPROVING FAIRNESS ON STUDENTS' OVERALL MARKS

We now proceed to design a new algorithm for the markers' group-wise reallocation. Let there be a total of M groups or classes of students, with the group size being approximately the same T,

and let there be a pool of N available markers denoted by 1, 2, …, N respectively. Some markers may not mark anything at all while some others may only mark some assessments. For the i-th group of students, their designated marker j for the k-th assessment will be denoted by $j=\rho(i,k)$. Although almost all of our later concepts or methodologies will be explained in just words, for the alternative more concise descriptions we nonetheless first introduce some notations for a few basic concepts or quantities.

Suppose K assessments have already been marked. For the k-th assessment, $1\le k\le K$, we assume that its weight in marks is w_k, and that each marker $j\in J\equiv\{1,2,\ldots,N\}$ was allocated the student groups $I_k^{(j)}\subseteq I\equiv\{1,2,\ldots,M\}$ and has already marked the allocated work for precisely the students of the set $S_k^{(j)}$, a total of $N_k^{(j)}\equiv|S_k^{(j)}|$ students who did submit the k-th assessment and belonged to the groups allocated to marker j. For any given marker $j\in J$, set $S_k^{(j)}$ could be an empty set $\varnothing$, or could consist of the students of one or several of the M student groups specified by $I_k^{(j)}$. Let $\{x_{k,s}^{(j)}\}$ for all $s\in S_k^{(j)}$ denote the marks for all the students marker j marked for the k-th assessment. We define the following variants of mark averages $\mu(j,k)$, $\nu(i)$, $\mu^*(j)$, and $\bar{\mu}(k)$ respectively by

$$\mu(j,k)=\left[\sum_{s\in S_k^{(j)}}x_{k,s}^{(j)}\right]\Big/N_k^{(j)},\ if\ N_k^{(j)}>0;\ \ \upsilon(i)=\sum_{k=1}^{K}\mu(\rho(i,k),k),\ i\in I;$$

$$\mu^*(j)=10\left[\sum_{k=1}^{K}N_k^{(j)}\mu(j,k)\right]\Big/\sum_{k=1}^{K}w_kN_k^{(j)},\bar{\mu}(k)=\left[\sum_{j\in J,s\in S_k^{(j)}}x_{k,s}^{(j)}\right]\Big/\sum_{j\in J}N_k^{(j)},\qquad(1)$$

where the mark average $\mu(j,k)$, conceptually being the average mark by the j-th marker on the k-th assessment, is defined in the above only when marker j did mark some student work, i.e. $N_k^{(j)}>0$. Otherwise, that is, if $N_k^{(j)}=0$, we simply set $\mu(j,k)=\bar{\mu}(k)$, which is defined as the average mark of all the submitted student work on the k-th assessment. Among *all* the assessments marker$j$ marked so far, the score percentage by marker$j$ is measured by $\mu^*(j)$ whose value will always be normalised to range from 0 to 10. For notational convenience, we here adopt 10 rather than 1 for $\mu^*(j)$ to represent 100%. A value 10 here would mean that corresponding marker would give full marks to every student work he marks. Let $\bar{\mu}(i,k)$ denote the average mark of the submitted work by the students in the i-th group on the k-th assessment. In the case of all groups of students having exactly the same number of students and all students submitted all their assessments for marking, then all quantities in (1) can be derived from the mark averages $\bar{\mu}(i,k)$ on each group for each $i\in I$ and k, see (2) later. In other words, there will be no need in this case to delve into the granularity of individual student marks once all the group averages are calculated.

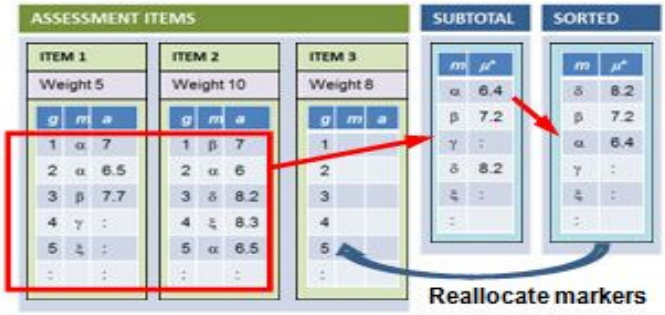

Fig. 1: Estimate and rank average marks for markers

With the preparations in the above, we are ready to reallocate the marking duties for the (K+1)-th assessment. For each group i, we will take $\nu(i)$ as the measurement of the total marking "biases"

the students in group i so far experienced and will take $\mu^*(j)$ as the prediction of the mark "percentage" for marker j for marking the next assessment. Our strategy is thus to allocate the most generous markers to the groups that experienced the most biases. The advantage of this strategy is that no marks are ever rescaled explicitly, but the total accumulated over several assessments will be objectively made fairer in that students are getting similar share of the toughest and the most lenient markers.

We now first sort student groups so that the groups of lower marks on average will appear earlier in the listing. More precisely, we set up a one-to-one mapping $\phi: I \to I$ such that whenever $a, b \in I$ and $a \le b$ we have $\nu(\phi(a)) \le \nu(\phi(b))$. Suppose for the next assessment item, each marker j is contracted to mark $M_j \ge 0$ student groups. We represent the marking duties by the set $D=\{ (j,n): j \in J, 1 \le n \le M_j$, and n is an integer$\}$, where each pairing (j,n) indicates an additional student group marker j needs to be assigned to mark. We now sort the marking duties so that their corresponding markers will have the more generous ones appear earlier in the listing. That is, we set up another one-to-one mapping $\psi: I \to D$ such that whenever $a, b \in I$ and $a \le b$ we have $\mu^*(a') \ge \mu^*(b')$ where $(a', x)=\psi(a)$ and $(b', y)=\psi(b)$. Then for the next assessment item, the i'-th student group will be assigned to marker j' such that $\phi(a)=i'$ and $(j', n)=\psi(a)$ for some $a \in I$, $j' \in J$, and integer $n \ge 1$. In plain words, we assign the most lenient marker to the student group which has been marked by the toughest markers in the previous assessments, see Fig. 1 for the estimation of the accumulated average marks for each marker and the subsequent ranking of these markers, where column g represents groups, column m represents markers, and column a represents mark averages. We note that the averages in Fig. 1 are always *represented* by a value between 0 and 10, with 10 indicating 100% of all the achievable marks, and this representation will also be adopted in the later tables or diagrams of this work. The complete reallocation procedure is summarised in the Markers Reallocation Algorithm (A1) framed out in the above.

A1: Markers Reallocation Algorithm

1. Assume there are M student groups of (approximately) the same size and N available markers, and marker j for each j is to be allocated $M_j \ge 0$ groups with $\sum_j M_j = M$.
2. Collect marks for all K marked items, and record which marker marked which student groups for which assessment (with k-th assessment weighing $w_k > 0$ marks) i.e. determine $\rho(i,k)$ for all groups $i \in I=\{1,\ldots,M\}$ and all items $k=1,\ldots,K$.
3. For each marker $j \in J=\{1,\ldots,N\}$ and item $k=1,\ldots,K$, calculate the marks mean $\mu(j,k)$ if j marked k-th item, otherwise assign to $\mu(j,k)$ the mean of all the student marks for the k-th item. Then $\forall i \in I$ set $\nu(i)=\sum_{k=1}^{K} \mu(\rho(i,k),k)$, and $\forall j \in J$ calculate $\mu^*(j)$ via (1) when satisfying $N^{(j)} \equiv \sum_{k=1}^{K} N_k^{(j)} > 0$.
4. If j marked no student work yet, set $\mu^*(j)$ to the default value μ^*, typically set to the weighted average $\mu^*=10\sum_{k=1}^{K} \bar{\mu}(k) \,/\, (\sum_{k=1}^{K} w_k)$.
5. Sort the mark averages $\{\nu(i): i \in I\}$ to an increasing order: construct an invertible mapping $\phi: I \to I$ such that the $\nu(\phi(i))$: $i=1,2,\ldots,M$, is in the increasing order.
6. Set the set D for the marking duties for the next item via $D=\{ (j,n): j \in J$, n is integer with $1 \le n \le M_j$ $\}$.
7. Sort the markers' averages $\{ \mu^*(\psi_1(i)): i \in I \}$ to a decreasing order: construct an invertible ψ: $I \to D$, $\psi(i) \equiv (\psi_1(i), \psi_2(i))$, such that $\mu^*(\psi_1(a)) \ge \mu^*(\psi_1(b))$, $\forall a, b \in I$ with $a \le b$.
8. Define $\rho(i, K+1) \equiv \psi_1(\phi^{-1}(i))$ which then maps each group $i' \in I$ to the reallocated marker $j'=\rho(i', K+1)$.

A2: Simplified Reallocation Algorithm

1. Same as A1(1), but group size identical.
2. Determine $\rho(i,k)$ as in A1(2) for all $i \in I$ and $k=1,\ldots,K$.
3. For all $i \in I$ and $k=1,\ldots,K$, calculate the mark average $\bar{\mu}(i,k)$ for the i-th group; then for all $j \in J$ calculate $\mu(j,k)$ from these $\bar{\mu}(i,k)$, if j marked the k-th item, via

$$\mu(j,k) = \left[\sum_{i \in I_k^{(j)}} \bar{\mu}(i,k)\right] / |I_k^{(j)}|. \quad (2)$$

Otherwise set $\mu(j,k)=\sum^{M}_{i=1} \bar{\mu}(i,k)/M$. Then $\forall i \in I$ set $\nu(i) \equiv \sum_{k=1}^{K} \mu(\rho(i,k),k)$, and $\forall j \in J$ calculate $\mu^*(j)$ via $\mu^*(j)= 10\sum_{k=1}^{K} w_k^{(j)} \mu(j,k)$ or via

$$\mu^*(j) = 10 \sum_{k=1}^{K} \sum_{i \in I_k^{(j)}} \bar{\mu}(i,k) w_k^{(j)} / |I_k^{(j)}|, \quad (3)$$

where $|I_k^{(j)}|$ is the number of groups j marked for the k-th assessment, and

$$w_k^{(j)} = |I_k^{(j)}| \,/ \sum_{k=1}^{K} w_k |I_k^{(j)}|. \quad (4)$$

4. Steps 4-8 are the same as A1(4-8).

In a perfect situation where all student groups are of absolute equal size and all students submit the assessments for marking, then the algorithm can be simplified to the Simplified Reallocation Algorithm (A2). We note that the simplified algorithm A2 in the above can be applied even if the group sizes are only approximately identical and a small portion of students may not submit some of their assessments, as the algorithm will nonetheless still reallocate markers for a fairer overall marks in general. The default μ^* in step 4 may also be set to $\mu^*=\sum^N_{j=1}\mu^*(j)M_j/\ (\sum^N_{j'=1}M_{j'})$. This way the markers who won't participate in the next marking task will be deemed irrelevant in predicting the average percentage in the next marking task.

4. IMPLEMENTATIONS

Group	Marker j	Marks average	Overall average
i	$\rho(i,K)$	$\bar{\mu}(j,K)$	$\nu(i)$
1	1	6	6.5
2	1	7	6.5
3	2	9	9
4	3	6	6

Fig. 2: Data on who marked which student groups and the mark averages with K=1.

j	$\mu(j,K)$	$\mu^*(j)$	M_j	$(j,1), .., (j,M_j)$	$\mu^*(j')$	j'
1	6.5	6.5	1	(1,1)	9	2
2	9	9	2	(2,1), (2,2)	7	4
3	6	6	0		7	5
4	7	7	1	(4,1)	6.5	1
5	7	7	0		6	3

Fig.3: Data on markers' averages with K=1. The marking profiles $\mu^*(j')$ are ordered in shaded columns.

Group i	Mark average $\nu(i)$	Sorted $\nu(i')$	i'
1	6.5	6	4
2	6.5	6.5	1
3	9	6.5	2
4	6	9	3

Fig.4: Sort group mark averages in shaded columns and define $\phi: I \to I$ by $i'=\phi(i)$ there. We then have $\nu(\phi(a))\le\nu(\phi(b))$ whenever $a,b\in I$ with $a\le b$.

To best illustrate the reallocation process, we now apply it to a simple case. We assume there are 4 student groups of 20 students each, and there are 5 available markers. Hence I={1,2,3,4} and J={1,2,3,4,5}. For the 1st assessment item, marker 1 is assigned to student group 1 and group 2; marker 2 is assigned to group 3 and marker 3 is assigned to group 4. Suppose that the average marks for the 1st assessment item for the 4 groups are 6, 7, 9, and 6 respectively, see Fig.2. For the

2nd assessment item, we have marker 1 available or contracted to mark just 1 group, marker 2 available for 2 groups, and marker 4 available for 1 group again, see the 5th column of Fig.3. We thus apply our above proposed algorithm manually to assign the markers to the suitable group/s so that the marks for the first 2 assessment items will overall be better balanced on the marking fairness in terms of the potential subjective biases.

First we calculate the average marks for each group from (1) or A2 with K=1, see the results for $\bar{\mu}(i,K)$ and $\nu(i)$ in Fig.2. Next we find the mark averages for each marker, and represent the marking of 4 student groups as the 4 pairs in the duty set D, see Fig.3. The value 7 in the column for $\mu(j,K)$ there is the average of the column for $\bar{\mu}(i,K)$, see (2) and the line below. Then we reorder the group indices i' to make tabulated group averages $\nu(i')$ increase downwards, see the shaded columns in Fig.4. We then reorder the marking duty pairs $(j',n')\in D$ so that the corresponding markers' averages $\mu^*(j')$ decrease downwards, see the middle 3 columns (shaded in light blue) in Fig.5.

$(j,n)\in D$	$\mu^*(j)$	$\mu^*(j')$	$(j',n')\in D$	j'	$\nu(i')$	i'
(1,1)	6.5	9	(2,1)	2	6	4
(2,1)	9	9	(2,2)	2	6.5	1
(2,2)	9	7	(4,1)	4	6.5	2
(4,1)	7	6.5	(1,1)	1	9	3

Fig.5: Sort the marking duties on the left to the 3 middle columns by markers' profiles, match with sorted group averages in dark shade on the right, then $i'\to j'$ maps the groups to the markers.

After appending the shaded columns on the right of Fig.4 to the right of Fig.5, we can complete the construction of $\phi(i)$ and $\psi(i)$ in Fig.6, and then read off the marker reallocation $\rho(i,K+1)$ directly from the table there. For reading convenience, we added sub-indices (a)-(f) to Fig.6 to indicate the sequential steps to arrive at (f) from (a). Hence, for the marking of the 2nd assessment item, marker 2 is assigned to group 1 and group 4, marker 1 to group 3 and marker 4 to group 2.

Rank	Group	Rank	Duties	Group	Marker
i	$\phi(i)$	i	$(j,n)=\psi(i)$	i'	j'
1	**4**	1	(**2**,1)	$1_{(a)}$	$2_{(f)}$
$2_{(c)}$	$1_{(b)}$	$2_{(d)}$	$(2_{(e)},2)$	2	4
3	2	3	(4,1)	3	1
4	3	4	(1,1)	**4**	2

Fig.6: Mappings for ϕ, ψ, and ρ. The shaded columns are the group to marker mapping $\rho(\cdot, K+1)$: $i'\to j'$.

As an example of applying this algorithm to a subject we delivered to a cohort of about 500 undergraduate students, the reallocated student groups can be directly shown to the markers via the web, see Fig.7, and can be accessed or downloaded by the markers in a single go.

We applied this reallocation strategy to an actual subject delivery in the 2nd semester of 2016 containing, among others, 2 assignments and a final exam. Before our measurement of the accumulated biases, we first remove the incomplete and irregular samples that often correspond to underdisciplined or extremely poor-performing students.

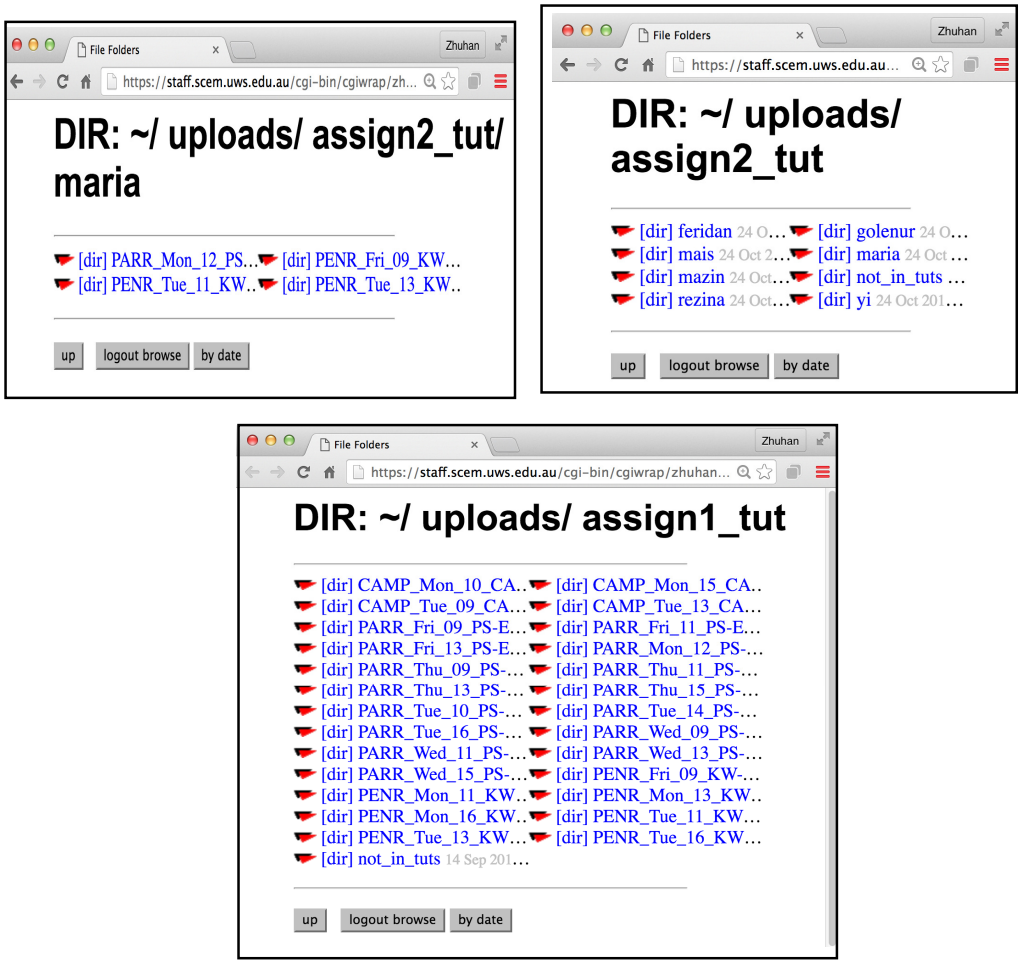

Fig. 7: Marking duties reallocated on web via the simplified A2 algorithm

Although we can treat all students as a single group, it is more illustrative to break students into several groups according to their final exam marks being G1:40-49.9%, G2:50-69.9%, and G3:70-100%. The largest of these groups has 121 students. We treated the final exam marks as more trustworthy as they are also more consistently marked, other marks will be compared against the exam marks. Since marks' consistency is largely relative, we will measure the consistency of the marks, $\{x_s\}$ and $\{y_s\}$ respectively, for two assessments by finding α and β such that such that $\{\alpha x_s+\beta\}$ is closest to $\{y_s\}$ in the sense of the least squares. The table in Fig.8 lists the least squares errors respectively for the 3 groups G1 (20 students), G2 (121 students) and G3 (87 students). We see that the accumulated errors have been reduced when assignments 1 and 2 are added together by compensating each other's marking leniencies.

Exam	40-49.9%	50-74.9%	75-100%
Item 1	2.5727	4.9335	6.3558
Item 2	2.7338	4.9253	6.4984
Items 1+2	2.5340	4.9181	6.2808

Fig.8: Estimated errors when comparing with the final exam marks.

While the algorithm in the above will try to objectively ensure that the marking of the student assessments is as fair as possible, we make sure that this message is also well disseminated to the students' mind so that they will also feel so themselves. We note that our proposed objective bias-balancing approach is the first of its kind as far as we know, although the first exploration of this kind was initiated in our previous work [15]. Our current work however will now be able to support the logistic fact that markers typically have to be assigned class-wise or tutorial group-wise, and the improvement on the fairer overall marks for the students such as those illustrated in Fig. 8 is on a similar scale as that exhibited in [15] when students have to be reassigned to different markers on the individual basis, therefore not suitable to beconducted class-wise, and also more involved statistical quantities such as variousstandard deviation of student marks will have to be made effective use of.

5. EXTENSION TO CROSS-COHORT BIAS-PROFILING

The advantage of utilising our proposed marker reallocation scheme is that it's effective, easy to implement, and applicable to all cases where students are grouped into tutorial classes and their assessment tasks are marked group-wise by the same marker. Nevertheless, there are still many different application scenarios that can induce further improvements on the fairness of the student marks. In what follows we will address in several subsections such specific application scenarios.

5.1. MARKING PROFILES AND ITERATIVE UPDATE

While our proposed marker reallocation scheme suffices for a single cohort of students, it's possible to retain certain essential form of marking statistics or profiles for potential future use. In this subsection, we will identify a set of crucial statistical features that can be easily extracted, archived and later made use of when needed. Suppose we have the marks of a total of K assessments marked by some of the markers in J. Then we may choose to retain only the following suite $\Im$ of data, also referred to as the cohort marking profile later,

$$\Im = \{\mu^*(j): j \in J, \nu(i): i \in I, T, w, W^{(j)}\}, \tag{5}$$

where $w=\sum^K_{k=1} w_k$, $W^{(j)}=\sum^K_{k=1} w_k \cdot N_k^{(j)}$, and the rest of notations are already defined in (1) or nearby. These data are more essential in the sense that an update on $\Im$ due to including the marks for the next assessment, i.e. the (K+1)-th assessment, can be calculated on the basis of the marks for the next assessment alone, without having to resort to any earlier data on the marks other than those of $\Im$ in (5). To this end, we just need to observe that $\mu(j,K+1)$ and $\bar{\mu}(K+1)$ in (1) can be calculated from the marks and marking allocations for the (K+1)-th assessment alone, while the accumulative $\mu^*(j)$, $\nu(i)$, w and $W^{(j)}$ can be shown to be updatable to $\mu'^*(j)$, $\nu'(i)$, w' and $W'^{(j)}$ via

$$\mu'^*(j) = \left[\mu^*(j) + 10\mu(j,K+1)\gamma_{K+1}^{(j)}\right] \Big/ \left(1+\gamma_{K+1}^{(j)}\right), \nu'(i) = \nu(i) + \mu(\rho(j,K+1),K+1),$$
$$\gamma_{K+1}^{(j)} = w_{K+1}N_{K+1}^{(j)}/W^{(j)},\ w' = w + w_{K+1},\ \ W'^{(j)} = W^{(j)} + w_{K+1}N_{K+1}^{(j)}. \tag{6}$$

This shows $\Im$ is a decent choice of the minimum data to keep for the marker reallocation algorithms. Once the marking of all the assessments is completed, we can treat the data in $\Im$ as the statistical profiling for the involved markers and archive them for the potential future use.

Since the quantity $W^{(j)}$ in (5) indicates the amount of the work j marked, we can normalize it into $W^{*(j)}$ via $W^{*(j)}= W^{(j)}/ \sum^N_{s=1} W^{(s)}$, so that $\sum^N_{j=1} W^{*(j)}=1$. Hence $W^{*(j)}$ may be treated as a form of *relative fidelity* on j's marking profile, as the more is $W^{*(j)}$, the more trustworthy is the profile $\mu^*(j)$. If all students submit all of the prescribed assessments, then all the $W^{*(j)}$ could also

completely determine all the $W^{(j)}$ because $\sum_{j=1}^{N} W^{(j)} = \sum_k w_k \sum_j N_k^{(j)} = wT_s$ and thus $W^{*(j)} = W^{(j)}/(wT_s)$, where T_s denotes the total number of students for the cohort. For the case of having M groups of students each with T students and the total marks weight is w, we then have $W^{(j)} = W^{*(j)}wTM$. The advantage of using the percentage oriented $W^{*(j)}$is that they add up to 100% and are more representative to their relative significance.

5.2. MERGE PROFILES ON TWO SEPARATE COHORTS

Suppose we currently have marking profiles $\Im'$ and $\Im''$ obtained from two cohorts of students. The cohort profile $\Im'$ corresponds to the primary cohort where notations will be mostly primed, and $\Im''$ corresponds to the secondary cohort where the notations will carry the double primes. If the set of common markers, $J \equiv J' \cap J''$, contains 2 or more markers, then it's possible to incorporate the marking profiles for some additional markers from J'' into that of J' for the primary cohort. Obviously the larger the cardinality $|J|$ is, the more meaningful this synchronization will be. We note that for a random variable of normal distribution, two values μ (mean) and σ (standard deviation) will suffice its determination, and one normal distribution can be easily rescaled to match another normal distribution. In fact, for each marker $j \in J$ if one denotes by $\mu^{(j)}$ and $\sigma^{(j)}$ respectively the mean and the standard deviation of the marks $\{ X_i^{(j)} \}_i$, $i=1, \ldots, L^{(j)}$, then for any $\mu^{\#}$ and positive $\sigma^{\#}$ the new marks $Y_i^{(j)}$ transformed by this rescaling

$$Y_i^{(j)} = \mu^{\#} + \left(X_i^{(j)} - \mu^{(j)}\right)\sigma^{\#}/\sigma^{(j)}, i = 1, \ldots, L^{(j)}, \qquad (7)$$

will be another normal distribution bearing$\mu^{\#}$ and $\sigma^{\#}$ as the new mean and standard deviation. In other words, (7) easily rescales one normal distribution into another that exhibits a more desired $\mu^{\#}$ and $\sigma^{\#}$.

Now we use the common markers in $J \equiv J' \cap J''$ as the pivot to rescale the profile $\Im''$ towards $\Im'$ as close as possible for the synchronization. For notational simplicity, we enlist J by $J=\{ s_1, s_2, \ldots, s_n \}$, and use x_j and y_j to represent $\mu''^*(s_j)$ and $\mu'^*(s_j)$ respectively. We now establish a mapping $y=P(x)$, with $P(x)=\sum_{j=0}^{m} a_j \cdot x^j$ and $m \geq 1$, such that the weighted error

$$f(a_0, \ldots, a_m) = \sum_{j=1}^{n} \lambda_j \left(y_j - P(y_j)\right)^2, \quad \lambda_j^2 = W'^{(j)} W''^{(j)}, \qquad (8)$$

is minimized. The non-negative weight λ_j is to ensure that its significance is proportional to the size of the past training data, or student marks to be more precise. In order to synchronize the profiles in $\Im''$ with those in $\Im'$, we need to rescale the profiles in $\Im''$ to fit as closely as possible within the perspective of $\Im'$ characterized by the profiles of the pivoting common markers. This is in fact a standard Least Squares problem, and the solution is $a = (A\Lambda A^T)^{-1} A\Lambda y, \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, where

$$a = \begin{bmatrix} a_0 \\ \vdots \\ a_m \end{bmatrix}, y = \begin{bmatrix} y_0 \\ \vdots \\ y_n \end{bmatrix}, A = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \vdots & x_n \\ \vdots & \vdots & \vdots & \vdots \\ x_1^m & x_2^m & \ldots & x_n^m \end{bmatrix}. \qquad (9)$$

We choose $m=2$ although m can be higher as long as $m \leq n$. However, a small m like 2 or 3 is more stable in general. We are now ready to merge some marking profiles from one cohort with that of another cohort. There are two slightly different ways merge them. The first case is to add the marking profiles for some of the markers from the secondary cohort $\Im''$ to the primary cohort. In this case, we in general expect the primary cohort $\Im'$ is still on-going and the data there is more

pertinent, and the merge is only required because new markers need to be introduced for the marking of the next assessment, typically due to some existing markers unexpectedly terminated their marking commitment. Hence we will retain the existing cohort profile specified by $\Im'$ as much as possible, and the merging is done only through adding extra markers in a selected subset of $J''\backslash J'$. The cohort profile $\Im'$ is then extended by including the following subset $\Im^e$ where

$$J = J' \cap J'', \Im^e = \left\{\mu'^*(j) = P\left(\mu''^*(j)\right), W'^{(j)} = \varphi^{(j)}, j \in J^e \subseteq J''\backslash J'\right\}, \text{ and}$$
$$\varphi^{(j)} = W'W^{*''}\left(\sum_{s\in J} W^{*'(s)}\right)\Big/\left(\sum_{s\in J} W^{*''(s)}\right), W' = \sum_{j\in J'} W'^{(j)} = \sum_{j\in J'}\sum_{k=1}^{|J'|} w'_k N_k^{'(j)}, \quad (10)$$

so that those extra markers in $J^e \subseteq J''\backslash J'$ can participate better in the reallocation for the next assessment because we will no longer assume these new comers are just the average of all the other known markers. We note that the choice of $\varphi^{(j)}$ in (10) is based on the proportion of $W^{*''(j)}$ against the proportion of the pivot elements in $J'\cap J''$ before it is translated back to the setting for the primary cohort. This is to preserve the relative significance of those additional markers from $J''\backslash J'$.

The second case is when none of the two cohorts is in any way primary. When a new cohort of students is to be marked for the first assessment, the default approach is to treat all the participating markers as having the same marking tendency or profile. This is of course not a problem on its own, but it can be further improved if we have one or more cohort marking profiles that contain some of the markers to be selected for the new cohort. If a cohort profile $\Im$ already contains all or enough potential markers to be selected, then we can select markers so that their marking profiles $\mu^*(j)$ are the closest to each other. If none of the existing cohort profile is sufficient to include all the candidate markers for the new cohort, then it's possible to combine two cohort marking profiles into a single virtual cohort profile so that it will contain more markers whose marking performances are made relatively comparable. This is our second case of merging the marking profiles for two cohorts. Recall that for the first case we have $\Im'\cup\Im^e$ where the marking profiles for markers in $J\equiv J'\cap J''$ are exactly those for the first (primary) cohort and are not affected by their behavior for the second cohort. Since there is no dominant cohort in the second case, we can take the average of $\mu'^*(j)$ and $\mu''^*(j)$ for those $j\in J$ and thus arrive at the following combined cohort profile

$$\Im = \left\{\mu'^*(j),\ W'^{(j)}, \mu''^*(j), W''^{(j)}:\ j \in J''\backslash J';\right.$$
$$\left.\left(\mu'^*(j)W'^{(j)} + \mu''^*(j)W''^{(j)}\right)\Big/\left(W'^{(j)} + W''^{(j)}\right),\ W'^{(j)} + W''^{(j)}:\ j \in J' \cap J''\right\}. \quad (11)$$

We note that the information on the existing average marks $\nu(i)$ will not make sense at this stage as no marking has been done yet for the new cohort and such non-existent $\nu(i)$ will come into existence only after the marking of the first assessment is completed.

5.3. NON-REGULAR VARIANTS OF APPLICATION SCENARIOS

The first thing a unit coordinator often needs to do in managing the marking of the student work is to select from a pool of broadly qualified but individually new candidate markers, and the general practice is to keep the same team of markers for the same unit for a whole semester as the marking contracts are typically semester based. It is of our no concern here in this work on how to select suitable markers from their resume or their past teaching experiences as we assume this will always be done properly anyway; rather we wish to make use of the existing marking profiles to help determine which candidates are likely to be more suitable in terms of their marking

consistency with the other markers. This is especially important if the number of the assessments or marking batches, which will be number of chances of reallocating the markers, is small or when the number is at the smallest value 1. In the very rare extreme case of having to have one marker to assess everything about each single student, our proposed reallocation scheme will not be suitable, and a remedy is of course to break the total assessment into a number of separate assessment items due at different dates. We note that which markers to select for the first assessment is essentially an extreme case of how to properly deal with bringing in new markers for a later assessment. The default for our proposed algorithm is simply assume every new marker is an "average" marker and there will be no differentiation among these newly added markers. However, in reality, the majority of those employed markers often already did some marking for the same unit in the past, and this means some comparison on their past marking profiles may be utilised to assist the selection of the markers. In this regard, we will select the marking profile for one or more past cohorts, and then select the markers there. If all candidate markers are within the group of markers for a past cohort, we select markers such that their marking profiles are closer to each other. If selecting enough candidate markers has to go across two or even more past cohorts, we may combine them together via (10) or (11), on the order of the most pertinent ones first, and them make the selection. We may also select only those markers of very similar profiles and choose some of the new markers to fill the remaining marking duties.

5.4. AN ILLUSTRATION EXAMPLE

In what follows we will present a simple example to illustrate the main concepts and procedures described in the previous subsections. Suppose we currently have at hands the marking profiles for two cohorts $\Im'$ and $\Im''$, with respectively the set of markers $J'=\{1,2,3,4\}$ and $J''=\{2,3,5,6\}$. For the current new cohort of students, all markers in $J'\cup J''$ have applied to mark for this new cohort, but we wish to select just 4 of them. For simplicity we assume (primed) first cohort has M=10 groups of students, each group has T=25 students, and the maximum achievable mark is w=20; the (double-primed) second cohort has also 10 groups of students, each group has 20 students, and the maximum achievable mark is w=15.

1st cohort $\Im'$			**2nd cohort $\Im''$**		
$j\in J'$	$\mu'^*(j)$	$W'^{*(j)}$	$j\in J''$	$\mu''^*(j)$	$W''^{*(j)}$
1	8.5	.1	2	5.7	.2
2	6.8	.4	3	6	.3
3	7	.3	5	5.8	.3
4	6.5	.2	6	7	.2

Fig.9: Marking profiles for 2 cohorts.

We also assume for presentational simplicity that all students will submit all of their prescribed assessments. The marking profiles are tabulated below, and $W'^{(j)}=wTM\cdot W'^{*(j)}=5000W'^{*(j)}$, $W''^{(j)}=3000W''^{*(j)}$. Since $J'\cap J''=\{2,3\}$ contains all the common markers, the function $f(a)=100\sqrt{15}g(a)$ for the least squares in (8) is $g(a_0,a_1)=\sqrt{8}\,[6.8-(a_0+5.7a_1)]^2+3[7-(a_0+6a_1)]^2$. Hence we obtain the mapping

$$\beta = P(\alpha) \equiv 3 + (2/3)\alpha \qquad (12)$$

via (9) with $\Lambda=\mathrm{diag}(\sqrt{8},3)$ and

$$A=\begin{bmatrix}1 & 1\\ 5.7 & 6\end{bmatrix}, y=\begin{bmatrix}6.8\\ 7\end{bmatrix}, a=\begin{bmatrix}3\\ 2/3\end{bmatrix}.$$

We are now ready to map the profiles for markers 5 and 6 in $\Im''$to $\Im'$via $\varphi^{(j)}$ in (10). In particular, we have from (12) $\mu'^*(5)=P(5.8)=20.6/3$, $\mu'^*(6)=P(7)=23/3$, and from (10) $W'^{*(5)}= W''^{*(5)}\times(W''^{*(2)}+W''^{*(3)})/ (W''^{*(2)}+W''^{*(3)}) \approx 0.42$ and $W'^{*(6)}\approx 0.28$. The marking profiles in $\Im'$ after the amalgamation then become those in Fig.10. With $W'^{(5)}=0.42W'$, $W'^{(6)}=0.28W'$ and $W'=20\times25\times10$, the profile amalgamation is completed although $W'^{(j)}$ are not really needed for our purposes.

$j\in J'\cup J''$	$\mu'^*(j)$	$W'^{*(j)}$	Renormalize $W'^{*(j)}$
1	8.5	.1	.0588
2	6.8	.4	.2353
3	7	.3	.1765
4	6.5	.2	.1176
5	6.867	.42	.2471
6	7.667	.28	.1647

Fig.10: Extended marking profiles.

We can now select the 4 required markers. We first calculate the weight average $\bar{\mu}^*= [\sum_{\text{all } j}\mu'^*(j)W'^{*(j)}] / [\sum_{\text{all } j}W'^{*(j)}]=7.0592$ from the above table, and then compare the $|\mu'^*(j) - \bar{\mu}^*|$. We find markers 3, 5, 2, and 4, in the order of preferences, are the ones with the 4 smallest differences, and hence we select these four as the markers for the new cohort. In theory, we could potentially expand the marking profiles further on in a similar way, but we expect that one such amalgamation should suffice in general. Once the markers have been selected for the marking of a new cohort, the amalgamated profiles will no longer be further needed as a more pertinent set of marking profiles will be generated entirely from the new marking tasks via our previous proposed algorithms.

In the case of some markers dropped out of the teaching team after having marked some but not all assessments, new markers need to be added to the marking team. If there are marking profiles for the new candidate markers from the previous cohorts, then the same amalgamation illustrated in the above can also be applied to this case without modifications. The amalgamated profiles will however be used only once, that is, for the immediate forthcoming assessment marking. Once the new markers' results are available, the synchronized profiles from a previous cohort will be replaced by the marking behavior of the current cohort. In other words, the impact of the cohort synchronization for the marking profiles will be largely just for one round of marker selection or reallocation as the later rounds will be based on the newly obtained marks. We finally note that any of above mentioned profile merging will be better than just assuming every unknown marker will be the average marker just like any vision is better than being blind, although quantitatively how much better will be left to our future investigation via systematic data simulation. At this stage, our actual use of this process is still semi-automatic in that we still largely follow the Fig.9 to Fig.10 steps here.

6. CONCLUSIONS

For a large cohort of students with a large team of inhomogeneous assessors, we can improve the fairness on the students' overall marks over the total collection of assessments by reallocating different markers to the students' different assessments so as to balance out the marking biases or inconsistency. Such a reallocation is best done at the granularity of tutorial groups, rather than via individual students. This way, the marking feedbacks can be better addressed in class, and will also be easier to be communicated to other markers pertinent to the individual groups. Moreover, the reallocation can be much more easily done since it conforms to the existing group formation. For candidate markers who served on the assessor teams in the past cohorts, we can synchronize

the marking profiles from a past cohort with those for the current cohort for selecting supplementary markers for the later assessments. We can also merge the marking profiles from two past cohorts so that a more consistent set of markers can be initially selected for the current new cohort of students.As for our future work, we will explore the impact and modification on our proposed strategies to achieve fairer student overall marks when peer assessments or other non-traditional assessment approaches are heavily involved, and we will also further investigate whether and how markers' prior marking profiles on different subjects can be better synchronised to predict their future marking inclinations.

REFERENCES

[1] Noesgaard, S.S., & Ørngreen, R. (2015). The effectiveness of e-learning: an explorative and integrative review of the definitions, methodologies and factors that promote e-learning effectiveness, Electronic Journal of e-Learning, 13(4), 278-290.

[2] Sun, P.C., Tsai, R.J., Finger, G., Chen, Y.Y., & Yeh, D. (2008), What drives a successful e-learning? An empirical investigation of the critical factors influencing learner satisfaction, Computers & Education, 50(4), 1183-1202.

[3] Fetaji, B., & Fetaji, M. (2009). E-learning indicators: a multi-dimensional model for planning and evaluating e-learning software solutions, Electronic Journal of e-Learning, 7(2), 1- 28.

[4] Pitt, E., & Norton, L. (2017). "Now that's the feedback I want!" Students' reactions to feedback on graded work and what they do with it, Assessment & Evaluation in Higher Education, 42(4), 499-516.

[5] Scott, S.,Webber, C. F., Lupart, J. L., & Scott, D. E. (2014). Fair and equitable assessment practices for all students, Assessment in Education : Principles, Policy & Practice, 21(1), 52-70.

[6] Admiraal, W., Huisman, B., & Pilli, P. (2015). Assessment in massive open online courses, Journal of e-Learning, 13(4), 207-216.

[7] Yalcin, M. A., Gardner, E., Anderson, L. B., Kirby-Straker, R., Wolvin, A. D., & Bederson, B. B. (2015). Analysis of consistency in large multi-section course using exploration of linked visual data summaries, PeerjPreprints, https://dx.doi.org/10.7287/ peerj.preprints.964v1.

[8] Reis, J., & Klotz, J. (2011). The road to loss of academic integrity is littered with SET: a hypothetical dilemma, Proceedings 5th Asia Pacific Conference on Educational Integrity, pp. 110-120.

[9] Kwan, K. P. (1999). How fair are student ratings in assessing the teaching performance of university teachers?, Assessment & Evaluation in Higher Education, 24(2), 181-195.

[10] Willey, K., & Gardner, A. (2010). Improving the standard and consistency of multi-tutor grading in large classes, ATN Assessment Conference, Sydney.

[11] Fermelis, J., Tucker, R., & Palmer, S. (2007). Online self and peer assessment in large, multi-campus, multi-cohort contexts, Ascilite 2007: ICT: Providing Choices for Learners and Learning, Singapore, pp. 271-281.

[12] Ashenafi, M. M. (2017). Peer-assessment in higher education – twenty-first century practices, challenges and the way forward, Assessment & Evaluation in Higher Education, 42(2), 226-251.

[13] Wilson, M. J., Diao, M. M., & Huang, L. (2015). "I'm not here to learn how to mark someone else's stuff": an investigation of an online peer-to-peer review workshop tool, Assessment & Evaluation in Higher Education, 40(1), 15-32.

[14] Barker, T. (2011). An automated individual feedback and marking system: an empirical study, Electronic Journal of e-Learning, 9(1), 1-14.

[15] Jiang, Z., & Huang, J. (2012). Bias Reduced Designation of Inhomogeneous Assessors on Repetitive Tasks in Large Numbers, International Journal of e-Education, e-Business, e-Management and e-Learning, 2(3), 176-182.

[16] Jiang, Z., & Huang, J. (2012). A fast and effective design and implementation of online programming drills, International Conference on Frontiers in Education: Computer Science and Computer Engineering, 314-320.

[17] Pepper, M. B. & Pathak, S. (2008). Classroom contribution: What do students perceive as fair assessment, Journal of Education for Business, 83(6), 360-367.

[18] Rojstaczer, S., & Healy, C. (2012). Where A Is Ordinary: The Evolution of American College and University Grading, 1940–2009, Teachers College Record, 114(7), 1-23.

AUTHORS

Zhuhan Jiang received the B.Sc. from Zhejiang University in 1982, and Ph.D. from the Victoria University of Manchester, Institute of Science and Technology, UK, in 1987. He is currently affiliated with University of Western Sydney, in the School of Computing, Engineering and Mathematics. His pertinent research interests include mathematical modelling and algorithms, web based security and applications, as well as image and video processing.

Jiansheng Huang received the B.E. and M.E. from Hefei University of Technology in 1982 and 1984 respectively, the MSc from The University of New South Wales in 1997 and the PhD from The National University of Singapore in 1999. Currently he is working with the School of Computing, Engineering and Mathematics. His relevant research interests include information systems and security, power system operation and protection.