Bias Reduced Designation of Inhomogeneous Assessors on Repetitive Tasks in Large Numbers

Zhuhan Jiang and Jiansheng Huang

Abstract—**Assessment consistency is not easy to maintain across many assessors for a unit of a large student population, particularly when a great many of those assessors are not regular staff. This work proposes an assessor reallocation approach, with some variants, to assign assessors to marking new assessment items for the different students based on the assessors' earlier marking statistics in comparison with that of the other assessors. This is to minimize the potential accumulation of marking discrepancies without having to resort to additional staff training which can often be impossible within the allowed time or budget frame. More specifically, we will first estimate the individual assessors' marking inclination or tendency, termed "bias" for simplicity, against the average for each particular assessment item, then profile each assessor by balancing such biases over a number of assessment items already marked, and subsequently predict for each student the potential bias he is likely to receive when marked by the different assessors. The proposed algorithm will finally select the assessor for the next assessment item so that it will lead to pro-rata the smallest difference with respect to the average of the accumulated total marking biases. This approach is objective and independent of the subjects being delivered, and can be readily applied, particularly in the context of e-learning or e-education, to any assessment tasks that involve multiple assessors in parallel over a number of assessment items.**

Index Terms—**Assessment consistency, balance assessors' biases, marks rescaling, universal algorithm.**

I. Introduction

Delivering a subject or unit to a huge cohort of students at a university, possibly across multiple campuses, is fundamentally more challenging than just to a small classroom. Many advances have been made in this regard including the use or development of various pertinent technologies [1]-[3]. For a cohort of several hundreds of students or more, a teaching team is almost always necessary to co-teach the whole unit with repeated deliveries at different timeslots and on different campuses. As the student demographics evolves from year to year, fluctuation in the cohort enrolment can deviate significantly from the most intelligent expectations. As a result of this, or simply because of the academic structure of a delivering school, some or even a great many of casual teaching staff members may have to be recruited to teach the same unit, often under a short notice. Regardless a teaching staff is permanent or casual, although academic qualifications and skills are in general suitably ensured, their teaching experiences can vary significantly and thus lead to potential inconsistency in both teaching and assessments. Such an inconsistency can be quite sizable if quite a few inexperienced new casual staff had to be employed to joint the teaching team. These problems that are inherent of large classes are also studied [4], [5] recently and it was found that the students' perception of difference in grade was not unfounded, and the problem was exacerbated by other factors such as the inconsistencies in the language markers use when providing the feedback. While effective and consistent assessment has always been actively studied, for such as the design [6] and for the group work [7], most of the studies on the assessment have been somewhat subjective and are also closely linked with the individual subjects. For instance, it was found [8] that the marking of a given assessment item tends to receive a better mark if the marking is immediately preceded by marking a poorer work of the same item. If the marking of the same work is preceded immediately by a better work then the marker tends to give a poorer mark. Moreover, individual assessors vary in their level of leniency or "biases", and the leniency of most assessors remains internally consistent throughout the local marking batches [9]. One exception to such subjective assessments is the so-called grading on the curve [10], [11] which determines the student grades according to the normal distribution of the marks. The main purpose of this work is to develop a subject-independent mechanism that can improve the overall marking consistency and fairness without having to delve into the performance of the individual assessors among a large number of disparate or inhomogeneous teaching team members. We note that even though our proposed work is possibly one of the very few, if not the first, of its kind in terms of an objective, universal and systematic algorithm in ensuring the marks consistency, these above mentioned conclusions by such as Spear and Lunz and O'Neil in fact also implicitly justify the consistency principles we use in devising our assessor reallocation scheme.

In a typical format of a tertiary unit of several hundreds of students across several campuses, we assume its delivery consists of repeated lectures, and repeated tutorial and/or practical classes of about 20 students per class, student assignments, an optional mid-semester test, and a final examination. Since it is feasible for the marking of mid-semester tests and final examinations to be maintained uniform and therefore fair across the student population, the largest marking discrepancy comes from the assessment of the student assignments as this often heavily depends on the casuals. The traditional approach to reducing the marking consistency relies on training the new staff, and supplying very specific marking criteria. However this may at times not

Manuscript received March 16, 2012; revised April 27, 2012.

Zhuhan Jiang and Jiansheng Huang are both with the School of Computing, Engineering and Mathematics, University of Western Sydney, Penrith, NSW 2751, Australia (e-mail: z.jiang{j.huang}@uws.edu.au).

be sufficient and the former may not be feasible. Moreover, subjective biases can be simply unavoidable when an assessment item is design based or even opinion based. Unless in the extreme cases where assessment errors are so obvious and prominent, it is simply not an option to ask a marker to do any form of reassessment. Marks rescaling for certain groups of students may seem as another valid choice, but it can be difficult to have it formally justified to all those involved. Our purpose here is thus to reallocate the marking tasks to different assessors for the different assessment items so that the inconsistency generic to the individuality of the assessors is well spread out and compensated over several items as much as possible.

This paper is organized as follows. First in section 2, we introduce a simple marks rescaling technique and then propose a framework to reassign the new marking duties according to the marks distribution of the earlier assessment items, so as to minimize the accumulated marking inconsistencies. Section 3 then investigates each individual marker's marking biases and measures their corresponding total bias accumulation. Section 4 subsequently proposes a major assessor reallocation algorithm, the Worse-Paired-First algorithm, along with several variant extensions. Section 5 is then dedicated to further justification of our proposed methodology via both the experimental simulations and an actual subject delivery. Finally section 6 gives a brief conclusion.

II. Marks Consistency and Fairness with Multiple Assessment Items

Marks inconsistency across different assessors is almost inevitable. For the staff collegiality it is never really a good practice to ask anyone to go back and refine parts of their earlier markings, unless blatant carelessness is observed for some of the casual staff. All these point to the need to develop a mechanism that can lead to more consistent marks through potential marks rescaling and/or shuffling the students among the assessors for the later assessment items. In this section, we will first give a rescaling scheme that can be applied directly to realign the marks that are already available from different assessors. We will then explore how to allocate later assessment items to different markers according to the existing marks statistics for their respective groups of students, so as to minimize the accumulation of the marking discrepancies or biases. For simplicity we may use *assignment* to refer to an assessment item of any form, and may use *tutor* to equivalently refer to an assessor or marker.

For a particular assignment, we assume that all the students who submitted the assignment will be organized into n groups each of which is assigned to a different marker. Each such group can be composed of the students of several tutorial classes although they will be typically selected from different tutorial classes for a "shuffling" of the markers. Let $T=\{1, 2, \ldots, k\}$ denote the set of all the tutors, i.e. the markers. Since it is generally accepted that student marks observe a normal distribution [10], [12], a simple rescaling described in (1) below can be applied to transform the marks consistently into those of a more desirable or anticipated distribution. Suppose for all the student marks M_s, s=1, …, K, the mean and the standard deviation are μ and σ respectively. Then for any target μ' and σ', the new marks M'_s rescaled from M_s via

$$M'_s = \mu' + (M_s - \mu)\sigma'/\sigma, \; s=1, \ldots, K \qquad (1)$$

will have μ' and σ' as the mean and the standard deviation respectively.

Suppose N assignments for each student have already been marked by a mixture of different tutors. Then for the next assessment item, the (N+1)-th assignment, which tutor should be assigned to mark which student's work so as to minimise the total bias that may be intrinsic to each individual tutor? Let $S=\{ s \}$ denote the set of all students, $T=\{ t \}$ with $t\neq 0$ denote the set of all the (marking) tutors, and $T^*=T\cup \{ 0 \}$ in which t=0 refers to a virtual assessor "responsible" for those who didn't submit the assignments. For each student $s\in S$ and each integer n with $1\leqslant n\leqslant N$, there exists a mapping ϕ_n such that $t=\phi_n(s)$ denotes that the n-th assignment for student s had been marked by tutor t if $t\in T$, or the assignment was not submitted at all if t=0. Our task is to determine an allocation mapping ϕ_{N+1}: $S\rightarrow T^*$ to best compensate the total biases already experienced in marking the first N assessment items.

Let $S_n=\{s\in S: \phi_n(s) \in T\}$. Then the set $S_n^{(t)}=\{s\in S_n: t=\phi_n(s)\}$ denotes all the students whose n-th assignment is marked by the t-th tutor. Hence for the n-th assignment, its marking details are completely determined by the S_n, ϕ_n, w_n, and ν_n, where w_n represents the positive weight of the assessment item and ν_n gives the grading percentage $x_{n,s}$ via $\nu_n(s)$ for each student $s\in S$, i.e. $x_{n,s}=\nu_n(s)$ and $0\leqslant x_{n,s}\leqslant 1$. The actual mark student s received for this n-th assignment is thus $w_n\cdot x_{n,s}$. We note however that in practice one may choose $x_{n,s}$ to be within the range of 0 to 10 because this wouldn't impact on our main scheme in this work but would be intuitively more meaningful.

For each student s, if tutor $t=\phi_n(s)$ has marked the student's n-th assignment, then the (percentage) mark $x_{n,s}$ will be denoted by $x_{n,s}^{(t)}$. This means the mark for the n-th assignment by the t-th tutor, $\{x_{n,s}^{(t)}\}$ for $s\in S_n^{(t)}$, has $K_n^{(t)} = |S_n^{(t)}|$ elements, where $|P|$ for any set P denotes the number of elements in the set. Hence $K_n=\Sigma_{t\in T} K_n^{(t)}$ is the total number of students who submitted the n-th assignment, and $K=\Sigma^{N}_{n=1} K_n$ is the total number of student submissions. For each tutor $t\in T$, we denote by $\mu_n^{(t)}$ and $\sigma_n^{(t)}$ respectively the mean and the standard deviation for the marks $\{ x_{n,s}^{(t)} \}$ for $s\in S_n^{(t)}$ given by tutor t for the n-th assignment. Likewise we denote by $\mu^{(t)}$ and $\sigma^{(t)}$ respectively the mean and the standard deviation for the $\{x_{n,s}^{(t)}\}$ over all the s and n with $s\in S_n^{(t)}$ and $1\leqslant n\leqslant N$. We also denote by μ_n and σ_n respectively the overall mean and standard deviation of the marks of the n-th assignment for the submitted students. For convenience we also set $\mu_n^{(0)}= \mu_n$ and $\sigma_n^{(0)}= \sigma_n$. We recall that μ and σ are the mean and the standard deviation of all the marks for the submitted N assignments. Since each tutor has his own marking tendency in such as marking leniently or harshly, we want to first establish the individual behaviour profile for each tutor in terms of his deviation or bias from the average by analysing their past marking statistics, and then assign tutors to the next assignment so that the accumulated biases are well-spread out and fair among all the students. The workflow of this strategy is depicted in Fig. 1, which will be gradually fulfilled later on.

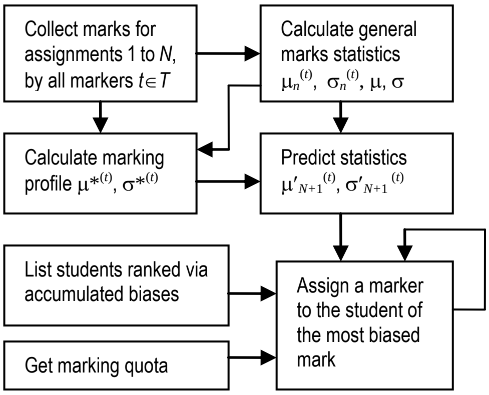
Collect marks for assignments 1 to N, by all markers $t\in T$
Calculate general marks statistics $\mu_n^{(t)}$, $\sigma_n^{(t)}$, μ, σ
Calculate marking profile $\mu^{*(t)}$, $\sigma^{*(t)}$
Predict statistics $\mu'^{(t)}_{N+1}$, $\sigma'^{(t)}_{N+1}$
List students ranked via accumulated biases
Assign a marker to the student of the most biased mark
Get marking quota

Fig. 1. Workflow to assign markers

Since the rescaling (1) is essentially a linear filter, and linear filters are known to exhibit rich properties, see e.g. [13], [14], we thus expect also a similarly rich collection of data relationships. For instance, if the data M_s are rescaled via (1) from μ and σ to μ' and σ', and then to μ'' and σ'', then it is the same as rescaling them directly from μ and σ to μ'' and σ'' in one step. Moreover, the transform $M_s \to M'_s$ in (1) is invertible whenever $\sigma\sigma' \neq 0$.

We are now ready to synthesise the statistics for the past marks. If the marks weight w_n for the n-th assignment is not equal to the weight $w_{n'}$ for the n'-th assignment, then we should also give different weight to the marking results. Hence we have

$$\begin{aligned} \mu &= \sum_{n=1}^{N} w_n K_n \mu_n \Big/ \sum_{n=1}^{N} w_n K_n \\ \sigma^2 &= \sum_n w_n K_n \left[\sigma_n{}^2 + (\mu-\mu_n)^2\right] \Big/ \sum_n w_n K_n \end{aligned} \qquad . \tag{2}$$

Since the smallest granularity for the statistics of the student marks is on the t-th group for the n-th assignment, we can derive them via

$$\mu_n^{(t)} = \sum_{s\in S_n^{(t)}} x_{n,s}^{(t)} / |S_n^{(t)}|, \quad \left[\sigma_n^{(t)}\right]^2 = \sum_{s\in S_n^{(t)}} \left(x_{n,s}^{(t)} - \mu_n^{(t)}\right)^2 / K_n^{(t)} \tag{3}$$

for all the t's and n's. In fact some other important statistics can also be presented in terms of these $\mu_n^{(t)}$ and $\sigma_n^{(t)}$ via

$$\begin{aligned} \mu^{(t)} &= \sum_{n=1}^{N} w_n K_n^{(t)} \mu_n^{(t)} \Big/ \left(\sum_{n=1}^{N} w_n K_n^{(t)}\right), \quad K_n^{(t)} = \left|S_n^{(t)}\right|, \\ \left[\sigma^{(t)}\right]^2 &= \sum_{n=1}^{N} w_n K_n^{(t)} \left[\left(\sigma_n^{(t)}\right)^2 + (\mu^{(t)} - \mu_n^{(t)})^2\right] \Big/ \left(\sum_{n=1}^{N} w_n K_n^{(t)}\right) \end{aligned} \tag{4}$$

for each marker t and

$$\begin{aligned} \mu &= \sum_{t\in T} K^{(t)} \mu^{(t)} \Big/ \left(\sum_{t\in T} K^{(t)}\right), \quad K^{(t)} = \sum_{n=1}^{N} w_n K_n^{(t)}, \\ \sigma^2 &= \sum_{t\in T} K^{(t)} \left[\left(\sigma^{(t)}\right)^2 + (\mu - \mu^{(t)})^2\right] \Big/ \left(\sum_{t\in T} K^{(t)}\right) \end{aligned} \tag{5}$$

for the overall. We note that we have also verified these formulas numerically.

III. Profiling Markers' Biases and Measuring the Accumulated Bias for a Student

We now examine how to effectively profile the markers' biases exhibited in the past marked N assignments. The term bias, however, should not be thought of as something that is necessarily wrong or even improper. It is here largely used to measure the extent of the inconsistency among the otherwise well-qualified markers. In other words, if each assignment had been marked by any single tutor for the whole cohort, then all the marks will be deemed appropriate, consistent and fair among all the students. Since it is in general not possible, even for a single assessment item, to have a single marker for a large cohort of students clustered on several campuses, the need to reallocate markers for the different assignments arises. The ever increasing trend of paperless electronic submissions by the students has also greatly facilitated this undertaking.

Without loss of generality, we will always assume that for any n the marks for the n-th assignment will not be identical for *all* the students, i.e. $\sigma_n \neq 0$, otherwise it wouldn't make sense to use such identical marks to measure the marking biases. Although one can simply collect all such marks and directly calculate the mean $\mu^{(t)}$ and the standard deviation $\sigma^{(t)}$ to measure their difference with the overall μ and σ for a particular tutor t, we should first take into consideration the inhomogeneity of the different assessment tasks. To envisage this, we first imagine that a single tutor marked the n-th assignment for the whole cohort in two halves. The mean and the standard deviation are calculated separately for these two halves. We can comfortably expect that they will be almost exactly the same, at least when the cohort is large enough. However, if a single tutor marked both the n-th assignment and the n'-th assignment for the whole cohort, we still have to expect that the statistics for these two different assignments will in general differ, regardless the size of the cohort.

In order to homogenize the statistics across the marks for the different assignments, we normalize the marks $x_{n,s}^{(t)}$ for each assignment to $x^*{}_{n,s}^{(t)}$ via

$$x^*{}_{n,s}^{(t)} = \bar{\mu} + (x_{n,s}^{(t)} - \mu_n)\, \bar{\sigma} / \sigma_n, \ \forall t\in T,\ s=1,..,K_n,\ n=1,..,N \tag{6}$$

against any base pair $\bar{\mu}$ and $\bar{\sigma}$. For the rescaled marks $\{\, x^*{}_{n,s}^{(t)} \,\}$ we will calculate the statistics $\mu^*{}_n^{(t)}$, $\sigma^*{}_n^{(t)}$, $\mu^{*(t)}$, $\sigma^{*(t)}$, $\mu^*{}_n$, $\sigma^*{}_n$, μ^*, and σ^* respectively in parallel to those without the *'s. These *-ed statistics constitute the bias profiling for all the tutors, and $\mu^* = \bar{\mu}$ and $\sigma^* = \bar{\sigma}$. More specifically, in our proposed tutor reallocation scheme, we will use $\{x_{n,s}^{(t)}\,\}$ to measure the biases accumulated over the marked assignments, use $\{\, x^*{}_{n,s}^{(t)} \,\}$ to predict the biases in the next assignment to be marked, and then reallocate the tutors so that the combined biases are shared evenly or pro-rata across all the students. In fact, $\mu^{*(t)}$ and $\sigma^{*(t)}$ can be explicitly represented by

$$\begin{aligned} \mu^{*(t)} &= \bar{\mu} + \bar{\sigma} \cdot \sum_{n=1}^{N} \xi_n^{(t)} \cdot (\mu_n^{(t)} - \mu_n) / \sigma_n, \\ \xi_n^{(t)} &= w_n K_n^{(t)} \Big/ \sum_{n'=1}^{N} w_{n'} K_{n'}^{(t)}, \\ [\sigma^{*(t)}]^2 &= \bar{\sigma}^2 \cdot \Big\{ \sum_{n=1}^{N} \xi_n^{(t)} [\sigma_n^{(t)} / \sigma_n]^2 + \sum_{n=1}^{N} \xi_n^{(t)} [(\mu_n^{(t)} - \mu_n) / \sigma_n]^2 \\ &\quad - [\sum_{n=1}^{N} \xi_n^{(t)} (\mu_n^{(t)} - \mu_n) / \sigma_n]^2 \Big\} . \end{aligned} \tag{7}$$

We note that even though we deliberately leave out the mathematical proofs here and elsewhere in this work for simplicity in order not to overstretch the coverage and complicate the main theme of this work unnecessarily with the derivational details, important formulae such as (7), (4) and (5) are nonetheless also verified numerically.

Let the statistics under tilde, such as $\tilde{\mu}^{(t)}$ and $\tilde{\sigma}^{(t)}$, denote those for the marks of the (N+1)-th assignment. Then we have

$$\tilde{\mu}^{(t)} = \tilde{\mu} + (\mu^{*(t)} - \bar{\mu})\tilde{\sigma}/\bar{\sigma}, \quad \tilde{\sigma}^{(t)} = \sigma^{*(t)} \cdot \tilde{\sigma}/\bar{\sigma} \qquad (8)$$

and we may also make use of $\mu^*=\bar{\mu}$, $\sigma^*=\bar{\sigma}$, and $\tilde{\mu}=\mu$, $\tilde{\sigma}=\sigma$ due to respectively (6) and the marks prediction. We note that the homogenization via (6) is essentially redundant if N=1, and this is because there are no multi-assignments that would need to be balanced out through the homogenization.

Since (7) allows us to predict the statistical behaviour or bias of each tutor for the (N+1)-th assignment, we now proceed to measure the accumulation of all the previous biases for each tutor. For simplicity we set $\bar{\mu}=\mu$ and $\bar{\sigma}=\sigma$. For a given student s, if tutor t' is assigned to mark the student's (N+1)-th assignment, then the total biases accumulated in the mean can be estimated as $D_\mu(s, t', w_{N+1})$ where

$$D_\mu(s,t',w') = \Delta_s\mu + w'(\mu^{*(t')} - \mu),$$
$$\Delta_s\mu = \sum_{n=1}^{N} w_n(\mu_n^{\phi_n(s)} - \mu_n), \qquad (9)$$

and $\mu_n^{\phi_n(s)} - \mu_n$ becomes 0 if student s didn't submit the n-th assignment. We note that D_μ in (9) simply represents the linear summation of all the previous N biases in mean (denoted by $\Delta_s\mu$) with the predicted bias in mean for the (N+1)-th assignment. As for the accumulated biases in the standard deviation, it can be similarly modelled by $D_\sigma(s, t', w_{N+1})$ via

$$D_\sigma(s, t', w') = |d^+ - d^-| + (2-\lambda)\cdot\min(d^+, d^-), \quad 0 \le \lambda \le 2, \qquad (10)$$

where $d^\pm(s,t',w')$ are respectively the sum of the positive and negative terms in

$$w'(\sigma^{*(t')} - \sigma)/\sigma + \Delta_s^+\sigma - \Delta_s^-\sigma, \qquad (11)$$

for $w'=w_{N+1}$, and $\Delta^\pm_s\sigma \ge 0$ are respectively the sum of the positive and negative terms in

$$\sum_{n=1}^{N} w_n(\sigma_n^{\phi_n(s)} - \sigma_n)/\sigma_n, \qquad (12)$$

where $\sigma_n^{\phi_n(s)}$ simply becomes σ_n if student s didn't submit the n-th assignment. We note that λ=0 corresponds to the extreme case of (direct sum of all the σ differences) $D_\sigma = d^+ + d^-$, λ=1 to D_σ=max ($d^\pm$), while λ=2 corresponds to the other extreme case of $D_\sigma = |d^+ - d^-|$. The denominators in (11) and (12) are to make the calculation percentage-wise. We also note that the case λ=0 in fact corresponds to the direct sum of all the accumulated deviation differences and the predicted deviation difference for the (N+1)-th assignment. The reason for using the model (10) is that the meaning of $\sigma' \equiv \sigma \pm \delta$ is not the same for the different signs even though $|\sigma'-\sigma|$ remains the same in both cases. To better gauge the closeness of σ and σ', we hence measured both the difference $|\sigma'-\sigma|$ and the overlap $\min(\sigma,\sigma')$.

IV. MARKER REALLOCATION ALGORITHMS

We are now ready to design the marker reallocation algorithm based on the statistics established in the previous sections. The main workflow for the reallocation is already depicted in Fig. 1. Some of the conditions governing the input and output can be summarised as the following

(C.1) $w_n > 0$, $\phi_n: S \to T^*$, $\nu_n: S \to \mathbb{R}^+$, n=1, ..., N; $w_{N+1} > 0$

(C.2) $S' \subset S$, $K'^{(t)} \ge 0$, $\sum_{t\in T} K'^{(t)} \ge |S'|$,

(C.3) $\phi(S') \subset T$, $\phi(S-S') \subset \{0\}$,

(C.4) $|\phi^{-1}(t)| \le K'^{(t)}$, $\forall t \in T$,

(C.5) $\exists\ 1 \ge \tau > 0$ such that $|\phi^{-1}(t)| \approx \tau \cdot K'^{(t)}$, $\forall t \in T$, (13)

where S' denotes the set of students whose new, i.e. (N+1)-th, assignments need to be allocated to suitable markers, and the allocation quota $K'^{(t)}$ denotes the maximum number of students whose new assignments could be contracted to be marked by the t-th tutor. The input to our algorithm should obviously include all the marks for the first N assignments and who actually marked which assignment for which student. This input can thus be symbolically represented by (C.1), and (C.2) implies that the students for the new assignment are already contained in the existing ones and the sum of the entire duty quota is sufficient to cover all the students. We note that providing the mapping ν_n is essentially the same as providing all the marks $\{x_{n,s}^{(t)}\}$. The expected output is simply a tutor-assigning mapping $\phi: S \to T^*$ such that (C.3)-(C.5) typically hold. In fact (C.5) is optional and implies that the total number of students assigned to each tutor should be proportional to their quota, if not all their quotas can be exactly met. (C.4) simply says the total number of students assigned to the t-th tutor should not exceed the quota $K'^{(t)}$. And (C.3) means the work submitted by the students in S' will be assigned tutors in T, and the rest of students didn't make the submissions. When one by one each student is being assigned a suitable tutor for the new assignment, students who suffered the most "biases" accumulated over the past N assignments should be given higher priority in finding the most suitable tutor so as to best compensate the past marking biases. For this purpose, we let U be an ordered list of the set S', sorted in the decreasing order of ($|\Delta_s\mu|$, $|\Delta_s\sigma|$), where $\Delta_s\mu$ is defined in (9) and

$$\Delta_s\sigma = |\Delta^+_s\sigma - \Delta^-_s\sigma| + (2-\lambda)\cdot\min(\Delta^+_s\sigma, \Delta^-_s\sigma) \qquad (14)$$

is defined along the same line as (10) and (11). In the case of two students having exactly the same ranking, the student with a higher total weight of the previously submitted assignments will be ahead of the other student. Hence the students who didn't submit any previous assignments will sit at the bottom of the list.

For a fairer scheme, we also randomise the ordering of the elements of U within the same band, i.e. the elements of the same $|\Delta_s\mu|$ and $|\Delta_s\sigma|$. Let $||s|| \equiv |\Delta_s\mu| + |\Delta_s\sigma|$ for all $s \in S$, we propose the following algorithm for the tutor reallocation.

WORST-PAIRED-FIRST ALLOCATION ALGORITHM:
Input: (C.1-C.2) in (13)
Output: (C.3-C.4) in (13)
Parameters:

ε: error tolerance for μ to compensate σ (default: 0)
λ: coupling constant in (10) and (14) (default: 0)
w': shorthand for w_{N+1}, i.e. $w'=w_{N+1}$

Variables:

U: list of students to be each allocated to a marker
U_t: set of students assigned to marker t
Q_t: adjusted marking quota for marker t
R_t: number of elements currently in U_t

Pre-processing:

- Set $Q_t \leftarrow K'^{(t)}$, and then reduce Q_t proportionally so that $\sum_{t\in T} Q_t = |S'|$.
- Round Q_t up or down to the nearby integer while maintaining $\sum_{t\in T} Q_t = |S'|$.

Algorithm:

```
i)    Initialisation: $U \leftarrow S'$; $U_t \leftarrow \varnothing$, $R_t \leftarrow 0$, $\forall\ t\in T$.
ii)   Calculate $\mu^{(t)}$, $\sigma^{(t)}$, $\mu$, $\sigma$, $\Delta_s\mu$, $\Delta_s\sigma$, $\forall\ t\in T$, via (4) etc.
iii)  Calculate $\mu^{*(t)}$ and $\sigma^{*(t)}$ via (7) for all $t\in T$, taking $\bar{\mu}=\mu$,
      $\bar{\sigma}=\sigma$, $\tilde{\mu}^{(t)}=\mu^{*(t)}$, and $\tilde{\sigma}^{(t)}=\sigma^{*(t)}$.
iv)   Sort $U$ in the decreasing order of $(|\Delta_s\mu|, |\Delta_s\sigma|)$, see (9) and (14).
v)    If $U=\varnothing$, go to step xii).
vi)   Let $s$ be the 1st element of $U$, set $\|s\|=|\Delta_s\mu|+|\Delta_s\sigma|$.
vii)  If $\|s\|=0$, randomly allocate all the students in $U$:
      a)  For all $t\in T$ such that $R_t < Q_t$, randomly pick $Q_t-R_t$
          students $V$ from $U$, assign them to tutor $t$, and then
          remove them from $U$, i.e. set $U_t = U_t \cup V$, $U=U-V$,
          and then set $R_t=Q_t$
      b)  Go back to step v).
viii) Otherwise (i.e. $\|s\|\neq 0$), find a $t'\in T$ such that
          $w'(\tilde{\mu}^{(t')} - \tilde{\mu})\times\mathrm{sign}(\Delta_s\mu)$,          (15)
      i.e. $w'(\mu^{*(t')} - \mu^*)\cdot\mathrm{sign}(\Delta_s\mu)$, is the smallest (i.e. most
      negative) among all those $t\in T$ with $R_t<Q_t$.
ix)   Find a $t''\in T$ such that $|D_\sigma(s,t'',w')|$ is the smallest among
      all those $t\in T$ with $R_t<Q_t$ and
          $|(\,|D_\mu(s,t',w')| - |D_\mu(s,t'',w')|\,)| \le \varepsilon$.          (16)
      If $t'\neq t''$, set $t'=t''$.
x)    Set $U'=\cap_{n=1}^{N} \{ s'\in U:\ \phi_n(s')=\phi(s) \}$.
      a)  If $|U'|\ge Q_{t'}-R_{t'}$, randomly pick $Q_{t'}-R_{t'}$ students $V$
          from $U'$, assign them to marker $t'$, and then remove
          them from $U$, i.e. set $U_{t'}=U_{t'}\cup V$, $U=U-V$, and then
          set $R_{t'}=Q_{t'}$.
      b)  If $|U'|< Q_{t'}-R_{t'}$, assign all students in $U'$ to marker $t'$,
          and then remove $U'$ from $U$, i.e. set $U_{t'}=U_{t'}\cup U'$,
          $U=U-U'$, and then set $R_{t'}=R_{t'}+|U'|$.
xi)   Go back to step v).
xii)  Each marker $t\in T$ is assigned to mark the new assignment
      for those students precisely contained in $U_t$.
```

There can also be variations on the above algorithm. For instance, the current algorithm is based on matching the student of the most positively biased total marks with the most negatively biased assessor for the (N+1)-th assignment, that is, via (15). However, for any given student, we can also search for a new marker that leads to the smallest accumulation of the biases instead of (15). Consequently we have a variant called

MANY-TO-NEXT ALLOCATION ALGORITHM:

(Note: The *Input*, *Output*, *Parameters*, and *Pre-processing* remain the same as in the WORST-PAIRED-FIRST algorithm)

```
i)-vii)  Same as in the WORST-PAIRED-FIRST algorithm.
viii)    If $\|s\|\neq 0$, find a $t'\in T$ such that $|D_\mu(s,t',w')|$ is the
         smallest among all those $t\in T$ with $R_t<Q_t$.
ix)-xii) Same as in the WORST-PAIRED-FIRST algorithm.
```

We note that U'in step x) contains students of the same status in that they have been assigned to the same tutor for the same assignment. Hence step x) allows such students to be essentially batch processed. We also note that step vii) a) makes good sense when those who didn't submit earlier assignments are equally likely to submit the latest assignment just like the others. However, the reality might be different in that students who failed to submit all the previous assignments are mostly the weak students and will most likely not submit the latest assignment either, since a later assignment is in general more challenging than the earlier ones. If this is the case, then step vii)a) would lead to those weak students' latest assignments being assigned to just one or two specific tutors who actually almost don't have to mark any of them at all since most of these students will probably not submit the latest assignment. We thus may optionally add to the above algorithms the following extra step vi′) between steps vi) and vii) for a fairer marking load among the tutors:

FAIRER MARKING LOAD STEP:

vi′): Let $S_0=\cap_{n=1}^{N} \{s\in S:\ \phi_n(s)=0\}$, i.e. S_0 is the set of students who didn't submit any of the previous assignments. For all $t\in T$, with T ordered in any particular sequence, if $|U_t| < Q_t$, randomly pick an $s\in S_0$, then set $U_t=U_t\cup\{s\}$, $S_0=S_0-\{s\}$ and $U=U-\{s\}$. Otherwise process the next $t\in T$ in the sequence. Repeat until $S_0=\varnothing$ or $U=\varnothing$.

It is perhaps worth noting that this load reduction was significant for some markers when we actually applied this step vi′) to units of large cohort of students.

V. Experiments and Evaluation

For the evaluation of our reallocation methods, we first evaluate the algorithms on the synthetically generated data, and then apply them to the deliveries of real subjects of large number of students, typical of service units or first year foundation units. For a given student, the marks for the different assignments are most likely correlated. If ψ_n is the true mark for the n-th of the first m assignments for a given student, then the true mark for the (m+1)-th assignment for that student is likely to be

$$\psi_{m+1} = (\alpha + \beta\Theta)\psi + \gamma\theta + \delta,$$
$$\psi = \left(\textstyle\sum_{1\le n\le m} w_n\cdot\psi_n\right) / \left(\textstyle\sum_{1\le n\le m} w_n\right), \tag{17}$$

where the mark ψ_{m+1} is to be truncated to its domain range if necessary, Θ and θ are two random numbers of unit Gaussian, and α, β, γ and δ are constants. This is based on the observation that a student's performance is generally consistent across several assignments.

In a real case of a subject delivery, one has only the actual marks for the marked assignments, and will never be able to get any hidden "true" marks. One way to make up for this is to compensate the assessors' biases on the actual marks and use these marks to substitute for the "true" marks. Hence for the actual marks $\{x_{n,s}^{(t)}\}_s$ for a given n, the rescaling towards the overall mean $x''_{n,s} = \mu_n + (x_{n,s}^{(t)} - \mu_n^{(t)})\,\sigma_n/\sigma_n^{(t)}$ will be treated as the approximate true marks. One can then *predict* the marks for the next assignment via (17), allowing a random variation of say 10% through the choice of the parameters there. After using $\mu_{n+1}^{(t)}$ and $\sigma_{n+1}^{(t)}$ to generate the "actual" marks for the (n+1)-th assignment, we can then evaluate the difference between the sum of the actual marks with the sum of the true marks, which will constitute our *predicted error*. We will show that these predicted errors will indeed decrease when the tutor reallocation is employed.

The complete reallocation algorithms and the simulation experiments are written in the form of a single program in PERL and the Box-Muller transform [15] is used to generate random normal distributions. Table I illustrates the typical reduction in the overall marking errors. Row A denotes the assessors 1-9, row K denotes the number of students a tutor will mark, $\Delta\mu$ denotes the difference of the individual mean with the given μ=7, and likewise for $\Delta\sigma$ for the given σ=2. The table shows that the errors for the use of "new" tutors are consistently smaller than those for the use of "same" tutors in all four error indicators: predicted, linear, squared, and least squares. In fact, we first generate a total of 416 true marks for the 1st assignment, utilizing repeated domain truncation and rescaling via (1). Then we allocate them to the tutors according to their quota and using their own marking mean and deviation to generate the actual marks the students will get, applying truncations again if necessary. We then similarly generate the true marks for the 2nd assignment, assuming that they are usually of up to 10% variation from those for the 1st assignment via (17). We first measure the accumulated errors if the same tutors mark the same students for both the assignments, then we measure the errors after the tutors are reallocated for the marking of the 2nd assignment. We note that different samples, with or without truncations, with or without the rescaling of individual marker's group of marks towards the prescribed mean and deviation, the results are very similar even though the actual numerical values will always differ due to the built-in intrinsic randomness.

TABLE I: SAME WEIGHTING ON BOTH ASSESSMENT ITEMS

A	1	2	3	4	5	6	7	8	9
K	41	53	93	87	23	16	41	39	23
$\Delta\mu$	-.19	.63	.40	-.27	-.37	-.77	-.41	.05	-.20
$\Delta\sigma$	.24	-.56	-.57	.03	.52	.83	-.04	.23	-.05

Error Type	Linear Error		Squared Error		Least Squares	
algorithm	same	new	same	new	same	new
many-to-next	.7920	.4397	1.090	.5087	1.082	.4880
prediction	.9936	.4019	1.310	.4828	1.292	.4587
worst-paired-1st	.7920	.4402	1.090	.5076	1.082	.4829
prediction	.9936	.3973	1.310	.4781	1.292	.4523

We observe that even though the predictions in Table 1 are based on the *estimated* true marks, they are very reasonable indicators on the error measurement. We note that if K marks $\{x_s\}$ are to approximate marks $\{y_s\}$, then $(\sum_s |x_s - y_s|^p/K)^{1/p}$ is the linear error for p=1 and is the squared error for p=2. If the approximation is in least squares, then the corresponding squared error is referred to as the least squares error. We also note that the data in Table I for such as A, K, μ, σ, $\Delta\mu$ and $\Delta\sigma$ are made to resemble an actual unit delivery to which Table Ⅳ is also related.

We now consider the case of 3 assignments with weight ratio 1:1:2, and with the same group of markers each assessing the same number of students as in Table I. We first randomly generate a total of 416 "true" student marks observing the normal distribution of μ=7 and σ=2, with possible truncation to the marks range and realignment towards the given μ and σ. These are then allocated to the tutors according to the **K** field in Table I sequentially. We again generate the corresponding "true" marks for assignment items 2 and 3 by assuming that the true marks for item 2 is a 10% perturbation of those for item 1, and the true marks for item 3 is a 10% perturbation of the synthesised marks for items 1 and 2, according to (17). We then compare the resulting marks difference between the true total marks and the actual total marks under different circumstances. We fired 1000 allocation simulations for the different cases and the results are summarised in Table II, where column name SSS indicates that the students each have the same marker of all the 3 items, SR indicates that students for the 2nd item are reassigned new markers via the reallocation algorithms and the comparison is just between 1st two items, and SRR indicates that students for the 2nd item are reassigned markers from the marks in the 1st item, and they are again reassigned new markers according to the marks for both item 1 and item 2. SRS and SSR denote that the markers have been reassigned for the 2nd and 3rd item respectively. The other controlling parameters are set to the default, and the shaded cells indicate the use of the WORST-PAIRED-FIRST method while the non-shaded data cells indicate the use of the MANY-TO-NEXT method.

In order to compare errors across multiple assessment items of different weights, the linear errors (linear), squared errors (L^2 norm) and least squares errors (LS) are all averaged against the total weight of the participating items. Table II convincingly shows that the marks errors will in general be cut into about half of what they would be when no reallocation is applied. Also, applying just the 2nd reallocation as in SSR has a very close effect of applying two consecutive reallocations as in SRR, because the weight of last item is more dominant and matches in total to the sum of the weights for items 1 and 2 together.

TABLE II: SIMULATION ON THREE ASSESSMENT ITEMS

Items	SRR	SSS	SRS	SR	SSR
Average (linear)	.2220	.3945	.2763	.2437	.2394
	.2274	.3945	.2716	.2375	.2391
Minimum (linear)	.1482	.3527	.2387	.1902	.1861
	.1597	.3527	.2344	.1820	.1861
Maximum (linear)	.2964	.4406	.3583	.3243	.3144
	.2917	.4406	.3478	.3255	.3184
Average (L^2 norm)	.3190	.5591	.3874	.3090	.3041
	.3254	.5591	.3809	.2989	.3036
Minimum (L^2 norm)	.2030	.5028	.3325	.2330	.2343
	.2109	.5028	.3292	.2284	.2343
Maximum (L^2 norm)	.4183	.6253	.5054	.4623	.4203
	.4424	.6253	.4963	.4497	.4583
Average (LS)	.2908	.5556	.3781	.2919	.2792
	.2967	.5556	.3708	.2793	.2784
Minimum (LS)	.1696	.4981	.3297	.2194	.2120
	.1790	.4981	.3251	.2148	.2120
Maximum	.3935	.6216	.4956	.4472	.4116

(LS)	.4005	.6216	.4927	.4422	.4439

Some desirable features such as reassigning markers as much as possible don't easily allow meaningful quantified measurement, since it is not really possible to determine for instance a different marker is worth how much in other form of error measurement. The preservation of distribution shapes (consistent Gaussians) via such as (16) in the algorithm is another similar example. However in the next simulation, we show that there does exist room for closer marks approximation when error tolerance ε in (16) in algorithm goes over the default value 0. For this purpose, we generated the "true" marks for item 1 similar to Table II, and removed the randomness of 10% for item 2. We allowed ε to vary between 0 and 1 at the step size .01. We found that there often exists such ε>0 which leads to better approximation (via the L^2 norm) to the true marks. This brute-force searching is done for 250 randomly generated cases, and we found for the best improvement the average ε is 0.1896 with the maximum (within [0, 1]) being 0.96. Other results are summarised in Table III.

TABLE Ⅲ: SIMULATION ON TWO ITEMS FOR CONSISTENT GAUSSIANS

Errors and Tolerance ε	Error Type	Linear Error	L^2 Norm	Least Squares
ε = 0	average	.2453	.3104	.2939
	maximum	.3496	.4558	.4360
Best case for 0 ≤ ε ≤ 1	average	.2370	.2988	.2833
	maximum	.3206	.4336	.4301
Improvement percentage	average	3.396%	3.764%	3.599%
	maximum	18.09%	15.51%	16.00%

We now apply our tutor reallocation algorithm to an actual unit consisting of, among others, 2 assignments and a final exam. Before measuring the accumulated errors, we first remove those incomplete samples (e.g. students missing one or more assessment items) and remove the extremely poor–performing students too because their marks are more likely to be "irregular" or even "non-truthful". Hence we conduct our evaluations on the students of complete records and in the groups of the exam marks exceeding 20, 25, and 35 respectively, out of the full mark 50. The largest of these groups has 290 students in total. Since marking consistency and fairness is largely in terms of the relative marks [11], it also makes sense to measure the errors in the least squares. In particular, when K marks $\{x_s\}$ approximate the marks $\{y_s\}$, we first find α and β such that $\{\alpha x_s+\beta\}$ is closest to $\{y_s\}$ in the Euclidean distance, and we could then also treat $\sum_s |\alpha x_s+\beta-y_s|/K$ as the average linear error. Table Ⅳ lists the least squares errors in the 2nd half, and lists the average linear errors in the 1st half (rescaled to the same level for comparison). It shows that the errors are being reduced across the board when assessment items 1 and 2 are added together, thus compensating each other's biases.

TABLE Ⅳ: ERRORS IN COMPARING WITH THE EXAM

Exam	20	25	35	20	25	35
Item 1	5.4516	4.5484	2.8278	5.8896	5.0287	2.8624
Item 2	4.3467	3.5108	2.7951	5.9888	4.6887	2.7951
Items 1+2	4.0604	3.3016	2.4051	5.7197	4.6313	2.7597

We finally note that the motivation to develop and analyse our tutor reallocation methodology comes from an earlier observation in teaching a database unit where over 20% marks difference were observed for apparently the same submission, partly due to a degree of subjective assessment on the design quality. This has thus led to the actual implementation of the above reallocation algorithms which have been by now routinely utilised for the 2 units of large student numbers for 3 consecutive semesters.

VI. CONCLUSIONS

We have proposed an assessor reallocation scheme, in the form of Worst-Paired-First algorithm and its variants, to objectively compensate the subjective marking biases by reallocating assessors to suitably different students for a new or next assessment item, by estimating all assessors' potential biases from the previously marked multiple assessment items. The simulations, along with the implementation on the actual university subjects, have demonstrated the convincing improvement by the scheme on the consistency and fairness of the assessments. This scheme offers for the 1st time a *universal* and tangible methodology to improve objectively the overall fairness on any sequence of student assessments.

REFERENCES

[1] F. Greyling, M. Kara, A. Makka, and S. Van Niekert, "IT worked for us: online strategies to facilitate learning in large (undergraduate) classes", *Electronic Journal of e-Learning*, vo. 6, pp.179-188, 2008.

[2] C. Scharber, S. Dexter, and E. Riedel, "Students' experiences with an automated essay scorer," *Journal of Technology, Learning and Assessment*, vol. 7, no. 1, 2008.

[3] Z. Jiang, X. Guo, N. Gangavarapu, and K. Khan, "Knowledge-based algorithms to optimise e-learning outcome," in *Proc. The 2009 International Conference on Frontiers in Education: Computer Science and Computer Engineering*, pp.247-253, 2009.

[4] K. Willey, and A. Gardner, "Difference in tutor grading in large classes: fact or fiction?" in *Proc.40th ASEE/IEEE Frontiers in Education Conference*, Washington DC, 2010.

[5] K. Willey, and A. Gardner, "Improving the standard and consistency of multi-tutor grading in large classes," in *Proc. ATN Assessment Conference*, Sydney, 2010.

[6] D. Trumpower, and T. Goldsmith, "Specificity of structural assessment of knowledge," *Journal of Technology, Learning, and Assessment*, vol. 8, no. 2, 2010.

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

[8] M. Spear, "The influence of contrast effects upon teachers' marks," *Educational Research*, vol. 39, no. 2, pp.229-33, 1997.

[9] M. E. Lunz, and T. R. O'Neil, "A longitudinal study of judge leniency and consistency," presented at Annual Educational Research Association, Chicago, pp. 24-28, March 1997.

[10] C. Wall, "Grading on the curve", *InCider*, vol. 5, pp.83-85, 1987.

[11] G. Kulick, and R. Wright, "The impact of grading on the curve: a simulation analysis", *International Journal for the Scholarship of Teaching and Learning*, vol. 2, no. 2, 2008.

[12] W. Hines, D. Montgomery, D. Goldsman, and C. Borror, *Probability and Statistics in Engineering*, 4th ed. USA:Wiley, 2003.

[13] Z. Jiang, and X. Guo, "A note on the extension of a family of biothorgonal Coifman wavelet systems," *The Australian and New Zealand Industrial and Applied Mathematics Journal*, vol. 46, pp.111-120, 2004.

[14] Z. Q. Huang, and Z. Jiang, "Image watermarking via sequencing wavelet filters," *Chinese Journal of Electronics*, vol. 17, no. 4, pp.649-654, 2008.

[15] G. Box, and M. Muller, "A note on the generation of random normal deviates," *Annals of Mathematical Statistics*, vol. 29, pp.610-611, 1958.