On Devising Objective e-Strategies to Improve Fairness in the Evaluation of Teaching Performance

Zhuhan Jiang*, Jiansheng Huang
School of Computing, Engineering and Mathematics, University of Western Sydney, Rydalmere, NSW 2116, Australia.

*Corresponding author. Email: z.jiang@uws.edu.au
Manuscript submitted September 20, 2015; accepted November 4, 2015.
doi: 10.17706/ijeeee.2015.5.4.249-256

Abstract: Evaluation of teaching performance is largely a modern phenomenon, and its systematic study is very much limited and scarce in comparison to its counterpart that investigates the students' learning performance. Since all evaluations on the teaching performance are meant to be utilized to enhance the attainment of the student goals, fairness in the evaluation by fair assessors can be crucial not only to the validity of the evaluation itself, but also to the ultimate fulfillment of the teaching and learning goals. This paper thus proposes a number of objective strategies, through the students' efforts and behavioral patterns, that can be applied uniformly to all students and determine the extent of their being a fair assessor in evaluating their instructors. Such strategies will encourage the instructors to concentrate on their core role of getting the students to learn better rather than bending the principles from time to time to please the students for the sake of getting a better student evaluation.

Key words: Fair teaching evaluation, objective performance evaluation, behavior-based adjustment.

1. Introduction

Modern technologies have greatly changed the way people live, work and study, and the traditional form of tertiary education has moved rapidly towards what is now called e-learning or blended learning [1], [2]. Not only standard software systems such as those course tools [3] are readily available for instructors to manage their course deliveries, but also a great many of additional software modules, commercially developed or in-house built, can be embedded or incorporated into those course tools to further enhance the functionalities and productivity [4]-[7]. Since the effectiveness of teaching and learning is generally attributed to the traditional teaching pedagogy, technological advances contribute more towards the efficiency on the delivery as well as its management. Other than the obvious effectiveness and efficiency of the knowledge delivery in e-learning [8], [9], one of the most important aspects in education lies in the fairness [10]-[12] in all the relevant assessments or performance evaluations. The evaluation in fact works in both ways: the evaluation on a student's performance is what we normally call marking or grading, which has been honed over the history of education, including our own previous work [5], while systematic evaluation on a teacher's performance is of a much more recent phenomenon linked to the trend of commercialization and is much less developed and relatively primitive. Since education nowadays is increasingly becoming like an industry, and competitions among the educators and students are everywhere, endless forms of performance evaluations have come into existence. While the composition of the items for evaluation and how they are evaluated can vary significantly [13], fairness is the essence that

needs to perpetuate all such evaluations.

A proper and fair performance evaluation on the teaching instructors should be designed to both evaluate and motivate them to achieve more and perhaps even outperform themselves. In this sense the design of the evaluation will play a crucial role in fulfilling its purpose, and a simplistic naïve design can be counterproductive at times. As a convincing example [14], a nine year long research in health care which explored linking doctors' income with their patients' satisfaction rating on the doctors has produced unexpected shocking results: patients in the surveyed group demonstrated a 238% increase in mortality and a 146% increase in morbidity with chronic medical conditions over the 10 year period. It was concluded that patient satisfaction appears to be directly related to increased mortality and morbidity, and an immediate halting on the relevant performance review program was called for. This case proves that a performance evaluation that hinges on pleasing the evaluators can be detrimental to the well being of even those pleased evaluators. This same naturally also applies the evaluation of the teachers by the students. Hence it is important that factors that could compromise the proper execution of the professional practice in exchange for an immediate better client satisfaction be reduced as much as possible, and our main goal in this work will be to seek such reductions for a better and fairer evaluation of the teaching performance. However, we need to note that fairness is not absolute, experience, culture, religions and many others can affect the judgment, and the widely accepted standard for the fairness may itself evolve with time even within the same sector of a society [15]. Since the perception of fairness can differ hugely when subjective assessments are involved, our purpose here is to devise measures as objective as possible to maximally improve the fairness. We will in fact also limit our consideration largely to the objective aspects, as well as to those that can be greatly assisted by the use of technologies. In particular, we will propose a number of strategies to alter the significance of a student's evaluation, in the form of a weight adjustment, to better reflect the truthfulness of the teaching evaluation by the students.

This work is organized as follows. We first in Section II explore how to adjust the weighting factor for a student's evaluation by his activity behaviors and show how to synthesize the results across different groups of instructors accordingly as indicated in Fig. 1 which will be further explained in the next section. Section III then demonstrates how the proposed strategies work through virtual implementations that will illustrate the inner working numerically. Finally section IV gives a conclusion.

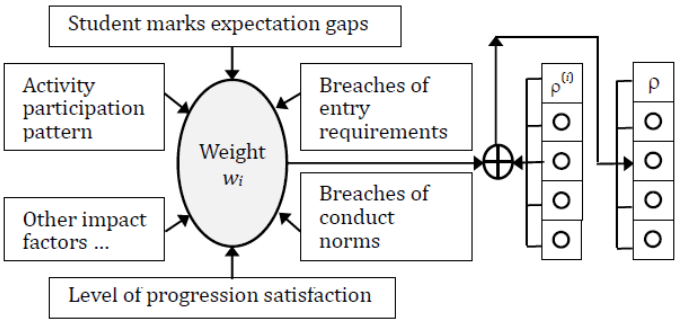

Fig. 1. Factors impacting the performance indicators.

2. Improve Fairness via Behavior-Based Adjustment and Result Synthesis

How a student evaluates an instructor's performance, is largely dependent on how much the student's expectations are fulfilled, and among all those expectations the grade expectation may weigh the most. However, a number of factors that solely depend on the individual students may implicitly impact on the achievement of their expectations. These factors could for instance include how well students follow the study guide, whether they abide by the rules without any form of misconduct, and whether proper pre-requisites are being met. Many of these study behaviors are well assumed by the instructors when they

try to optimize their deliveries, and any unilateral ignorance on the behavior protocols by the students could eventually lead to the lack of satisfied expectations, often to the no faults of the instructors. We hence propose to introduce an objective mechanism that would reduce the importance of a student's evaluation if he or she breaches the commitment or protocols expected of a student. Fig. 1 depicts our main design strategies, where different impact factors are taken into consideration to improve the fairness on any performance indicator $\rho^{(i)}$ for each student i, and the $\rho^{(i)}$'s for all the students are then synthesized into the final valueρ.

2.1. Evaluation Adjustment Based on Students' Behaviors

The crucial aspect in our proposed system is that we will add appropriate weighting to the evaluation results according to a student's true involvement with the activities or subjects we are trying to evaluate. The followings are some of the factors we propose to be taken into consideration:

1) If a student never attends the classes, then the student is deemed not doing what he should be apart from the possible exceptions we will address below shortly, and will therefore be reduced in the weighting of his feedback opinion. This will motivate students to increase the class attendance or its equivalent activities. However, for some subjects or types of studies, students may be waived the compulsory class attendance as long as their accumulated assessment marks go beyond a certain threshold such as 85%. We may thus represent such a weight adjustment $w \rightarrow w'$ mathematically as $w' = \lambda 1 \cdot w$, with

$$\lambda_1 = \min\left\{1, \alpha + (1-\alpha)\left[\frac{n}{N} + \max\left(0, \frac{m/M - \beta}{\gamma - \beta}\right)\right]\right\} \quad (1)$$

where w is the current feedback weighting, N denotes the total number of classes to attend, n the actual number of classes attended, M the total mark achievable thus far, m the total mark obtained, and α, β and γ ($\gamma > \beta$) are three chosen parameters. We note that w is always a number between 0 and 1 (100%). If one chooses for instance α=.5, β=.8 and γ=.9, then it means everyone will have at least α=50% on the feedback weight, a student will have the full feedback weight if he has achieved, say 90%, on the total assessment marks regardless of his attendance, a total mark percentage m/M being less than β=80% of the total achievable will not contribute towards the additional weight, and any mark in between will contribute pro-rata.

2) If a student received academic penalties for some form of misconduct, then the right to evaluate teaching performance should be revoked, i.e. $w' = \lambda 2 \cdot w$, where $\lambda 2$=0 whenever there is an established misconduct on the student's part. This will thus allow the teaching staff to enforce school or class regulations without worrying about the potential feedback retribution.

3) If a student was enrolled into a subject without fulfilling the full prerequisites first, in other words he is in some way "accidentally" enrolled, then we might also slightly deduct the student's evaluation weight from w to w' via $w' = \lambda 3 \cdot w$ with

$$\lambda_3 = \max\{0, 1 - \delta \cdot P\} \quad (2)$$

where P denotes the total number of unfulfilled pre-requisites, and $\delta > 0$ is a chosen impact parameter. This formula implies the more pre-requisites a student lacks, the less evaluation weight the student will be given. As an example, if one sets δ=0.2, then missing a pre-requisite will reduce by 20% the student's evaluation weighting.

4) We can also ask students to give their expected marks for their assessment tasks before being marked,

even though these self-assessed marks may be kept away from the markers. Suppose there are a total of K assessment items for the students that have been completed and marked when a teaching performance evaluation is conducted, and the i-th assessment item carries the weight ω_i towards the final grade, and E_i and T_i are respectively the expected and true marks for the i-th item. Then the total over-anticipation on the marks can be measured by

$$\Delta = \left(\sum_i \omega_i \Theta\left(E_i - T_i\right)\left|E_i - T_i\right|/\left|T_i\right|\right)/\left(\sum_i \omega_i\right), \tag{3}$$

where $\Theta(x)=1$ if $x>0$ and $=0$ if otherwise. The use of Θ function implies that a student's under-estimation of his assessment marks will not affect his weighting as it would not incur biases. The resulting weight adjustment then becomes $w' = \lambda_4 \cdot w$, in which

$$\lambda_4 = \min\left\{1,\ \xi + (1-\xi)\left[\max\left(0, 1-\tau(\Delta-\eta)\Theta\left(\Delta-\eta\right)\right)\right]\right\}, \quad 0 < \xi, \eta \le 1,\ \tau > 0, \tag{4}$$

where ξ is the minimum weight, i.e. $\lambda 4 \ge \xi$, η is the tolerance related to a legitimate amount of expectation error, and τ is the impact magnitude. We note that if E_i doesn't exist for a particular i, then we can treat it as (i) $E_i = M_i$ or (ii) $E_i = T_i$ or (iii) just take that i out of consideration, where M_i is the maximum mark for the i-th assessment item. The defaulting strategy (i) is generally speaking more ideal, although strategy (iii) is also quite reasonable as long as each student does at least one mark estimation himself.

There can be many other student behaviours that are deemed desirable or undesirable for the proper and effective learning process, and we can similarly adjust the student feedback weighting to promote those positive behaviours while discouraging the negative ones. In other words, students who "follow the rules" will have their feedback heard in the loudest voice. Hence there can be further weight adjustment factors λi and the final weight for a given student will thus be $w=\lambda=\lambda 1 \cdot \lambda 2 \cdots \lambda t$, the product of all the adjustment factors. The controlling parameters α to δ and etc. may be set intuitively, or with a specific purpose. They may also be dynamically adjusted to meet certain fairness criteria. For instance, one may dynamically optimize these parameters so that a dominant portion, say 2/3, forms the shortest cluster diameter. This however has gone beyond the scope of the current work. In the environment of e-learning, each student can check their current "weight" w according to their intermediate results or activities. This awareness may also positively motivate the students to follow more closely the general rules and guides set out by the instructors.

2.2. Synthesize and Unify Performance Indicators over Multiple Groups of Instructors

We have so far considered only how to adjust the weight for a given student's evaluation of his instructor's performance. Next we consider how to synthesise the evaluation results from all the students. For a teaching performance, there may be a number of indicators, and we use ρ to denote a typical such performance indicator. Suppose there are S students participating in the performance evaluation of their instructors, $\rho^{(j)}$ is the value the j-th student gives for the indicator ρ, and $w^{(j)}$ is the final weight for the j-th student, then the final value for the indicator should read

$$\rho = \left(\sum_{j\in S} w^{(j)}\rho^{(j)}\right)/\left(\sum_{j\in S}\rho^{(j)}\right) \tag{5}$$

It is obvious that our proposed approach is well suited to any form of online teaching evaluation. The existing online Student Evaluation on Unit system at our institute gives each student a period of 6 weeks or so to complete the evaluation of any delivered subject. Students will be anonymous to the teaching staff.

However, the fact that the online feedback system requires to authenticate each student first implies that feedback can be logged and the feedback response is traceable back to the individual student by the institute although teaching staff will not have such abilities. This underlying logging is nonetheless beneficial in that it allows the hosting institute to conduct more correlated analysis and the students are indirectly held responsible for what they say. For instance, no students will ever be able to put outrageously false stories in their comments, and the student responses must also conform to the acceptable standard and ethics. In a way, this is already a step towards making the student feedback on the teaching fairer, compared with the completely anonymous paper-based on the spot evaluations. Our proposed approach can be easily added to such an existing system or the like to better address the fairness of the teaching evaluation. With the online logging facilities, our feedback system will go a long way on improving fairness of the system in comparison with the traditional paper-based version, and it is also possible as our future work to train the parameters α to δ and etc. to optimize their values.

In a completely different perspective, there can also be potential inherent biases towards certain groups of instructors due to such as their genders, races and social grouping. For instance, it was found [16] that female college teachers are more likely to get better student evaluation than their male counterparts. This is possibly also akin to being unfair if one expects the physical strength of an average woman to match that of an average man. Since it's not easy to precisely match a rating for one group to another rating for a different group, one may ideally conduct such comparisons among a similar group of an acceptable statistical size. If there needs to be a rating applicable across different instructor groups, then we can employ the following rescaling method for this purpose.

Suppose we are considering a particular single indicator ρ for the teaching performance, and assume that G={1, 2, ..., g} denotes the set of all the instructors who take part in their performance evaluation by the students, and μ_k and σ_k denote respectively the mean and the standard deviation of the ρ values for the instructors of the k-th group and all of these groups are mutually disjoint. We also assume that indicator values will observe a normal distribution. Then one way of consistently rescaling the evaluation results in the form of the indicator values is to transform them via

$$\rho'_j = \mu' + (\rho_j - \mu_k)\sigma'/\sigma_k, \qquad j \in G, \tag{6}$$

where instructor j belongs to the k-th group. The advantage of this mapping, similar to the case adopted in [5], is that for each group of instructors the mean and the standard deviation of the transformed indicators ρ'_j will all become just μ' and σ' respectively. If we choose μ' and σ' to be just the mean and the standard deviation of the original set of values $\{\rho_j\}_{j\in G}$, then all groups will be aligned to the same mean μ and the same standard deviation σ. This way, the performance indicators become more comparable across different groups. We note that this approach is valid only when each group is large enough to reasonably well observe the statistical properties. Hence forming very small groups of special interests and conducting the analysis this way will not be very meaningful.

3. Virtual Implementations

Since a student has to login before being allowed to conduct a teaching evaluation, the system can easily log the student details and link them with the evaluation data. Moreover, such logging will be able to allow a third party to link the evaluation data with the student records, such as whether the student was given a misconduct during the semester or how many mandatory classes or activities the student has missed out, without the awareness of who the evaluator is by the relevant instructors. Due to the logistic difficulties to clear the necessary technical, ethical or confidentiality hurdles in a short space of time, the implementation has yet not been done on the actual subjects. Instead we will consider a virtual implementation at this stage

by looking into different potential student cases and examine their impact towards producing the final evaluation results.

We first illustrate the case with formula (1). In Table 1 we set α=.5, β=.65, γ=.9, denote by a_i (with $0 \le a_i \le A_i$) the attendance of i-th activity or class, denote respectively by M_j and m_j to the maximally achievable mark and the actual mark obtained for the j-th assessment item, and denote by $S^{(k)}$ to k-th student. Hence $n=\sum_i a_i$, $m=\sum_i m_i$, $N=\sum_i A_i$, and $M=\sum_i M_i$. We note that (i) student $S^{(1)}$ performed exceedingly well with his assessments and hence his weight on the performance evaluation is not downgraded despite of his frequent absence of class attendance; (ii) student $S^{(2)}$ made reasonable attendance efforts with reasonable marks hence his weight is not downgraded by (1) either; (iii & iv) the students there didn't perform well in assessments and didn't have good attendance, hence their weight is downgraded; and (v) student $S^{(5)}$ made perfect attendance and hence got the full weight even though his assessment marks are very poor. We note that if we made the submission of the assessments compulsory, in other words the submissions will be treated as a form of "attendance", then this student $S^{(5)}$ wouldn't be given the full weight 1 because his "attendance" wouldn't be full under this new circumstance.

Table 1. Attendance and Marks Impact on λ_1

	Attendance					Assessment Marks				Results	
	a_1	a_2	a_3	...	n/N	m_1	m_2	...	m/M	λ_1	Note
$S^{(1)}$	.5	0	0		2/3	9	14		.92	1	(i)
$S^{(2)}$	1	.5	1		5/6	7.5	10.5		.72	1	(ii)
$S^{(3)}$	1	0	0		1/3	5	7		.48	.67	(iii)
$S^{(4)}$	.5	0	0		1/6	3	5		.32	.58	(iv)
$S^{(5)}$	1	1	1		1/6	1	0		.04	1	(v)
Max	1	1	1	A_i	N/A	10	15	M_i	N/A		

Table 2. Weight Impact by Student Expectation on Marks or Grades

	T_1	E_1	T_2	E_2	T_3	E_3	ε	Δ	λ_4	Note
$S^{(1)}$	5	7	10	14	5	7	.2	.4	.8	(i′)
$S^{(2)}$	8.5	9.5	16	18	7	8	.1	.1276	.9712	(ii′)
$S^{(3)}$	6.5	7	12	13	7	7.5	.05	.0788	.9956	(iii′)
$S^{(4)}$	5	5.3	9	9.6	6	6.3	.03	.0608	1	(iv′)
Weight	ω_1	10	ω_2	20	ω_3	10		ξ=.5, η=0.07, τ=1		

Since the determination of λ_2 and λ_3 is rather straightforward, we will move to directly experiment with λ_4 instead. By choosing η=0.07, we will not penalize any students who overestimate their achievements by no more than about 3% of the maximum total mark, and by reducing τ from the current value 1, we can further soften the penalty on the students' unrealistic grade or marks expectations which often lead to negatively biased evaluation of the instructors' teaching performance. In Table 2, we denote by ε the overestimated extra mark over the maximum total mark, and we note that (i') student $S^{(1)}$ overestimates his performance by 20%, and thus get a larger reduction on his weighting; (ii′-iii′) the students there with overestimations at 10% and 5% respectively will get a very small weight deduction; and (iv′) student $S^{(4)}$ will have no weight deduction because his overestimation is within the acceptable 3%.

For the synthesis of all the students' evaluations on the teaching performance, we illustrate the case of 5 students in Table 3, where the overall ρ=4.5 for the regular average is improved to ρ=4.56 under the proposed synthesis scheme.

Table 3. Synthesize the Evaluation Results from Multiple Students

	$S^{(1)}$	$S^{(2)}$	$S^{(3)}$	$S^{(4)}$	$S^{(5)}$	Average	Final ρ
j	1	2	3	4	5		
$w^{(j)}$	0.9	1	0.8	0.6	0.7	$\sum_j w^{(j)}\rho^{(j)}) / \sum_j \rho^{(j)}$	4.56
$\rho^{(j)}$	5	4.8	4.5	4	4.2	$(\sum_j \rho^{(j)})/N$, where $N=\sum_j 1$	4.5

As our final experiment, we move to consider the use of the rescaling (6) to offset the potential generic differences across different groups of instructors. Suppose we partition all the instructors into 2 groups G_1 and G_2, for instance male and female groups, with 4 members in each group undertaking the performance evaluation. The sizes of the groups can be arbitrary as long as they are reasonably large to exhibit good enough their statistical features. Table 4 first calculates the means and standard deviations on each group individually, and then calculates the corresponding rescaled values. Under this scheme, the original performance indicators shaded in grey will be transformed to the final adjusted indicators shaded in cyan. These adjusted performance indicators will now make the instructors more comparable across those different groups.

Table 4. Rescale Indicators to Reduce Group Biases

	1	2	3	4	μ	σ	$G_1 + G_2$	
G_1	4.5	5	4.5	4	4.5	.3536	μ	4.725
G_2	5	5.5	4.5	4.8	4.95	.4640	σ	.4235
G_1	4.725	5.324	4.725	4.126	4.725	.4235	μ′	4.725
G_2	4.783	5.365	4.201	4.550	4.725	.4235	σ′	.4235

We finally note that these virtual implementations and demonstrations can be easily implemented on a real-life performance evaluation system, and can be automated if the surveys are to be done online. The pre-selected parameters can either be manually adjusted to reflect a managerial decision, or get trained over a number of evaluations to optimize certain specific goals.

4. Conclusion

It is very important that teaching instructors should not be adversely affected to a great extent in regard to their performance evaluation by the students, by the instructors' needs to uphold the right educational principles and rules. By adjusting a student's weight on the evaluation of his instructors to reflect his efforts and behavioural patterns through a number of objective measures, we are able to promote better learning practice and teaching effectiveness, and achieve fairer evaluation on the teaching performance at the same time. Moreover, we are also able to properly rescale the performance indicators to make teaching instructors of generically inhomogeneous groups comparable through the common unified performance indicators.

References

[1] 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.

[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] Greyling, F., Kara, M., Makka A., & van Niekerk, S. (2008). It worked for us: Online strategies to facilitate learning in large (undergraduate) classes. *Electronic Journal of e-Learning, 6(3)*, 179-188.

[4] Barker, T. (2011). An automated individual feedback and marking system: An empirical study,

Electronic Journal of e-Learning, 9(1), 1-14.

[5] 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.

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

[7] Jiang, Z., & Guo, X. (2012). A miniature e-learning portal of full teaching and learning capacity. *Proceedings of the International Conference on Consumer Electronics, Communications and Networks* (pp. 3027-3030).

[8] 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.

[9] Wang, T. H. (2010). Web-based dynamic assessment: Taking assessment as teaching and learning strategy for improving students' e-learning effectiveness, *Computers & Education, 54(4)*, 1157-1166.

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

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

[12] Slade, P., & McConville, C. (2006). Student evaluation of teaching. *International Journal for Educational Integrity*, *2(2)*, 43-59.

[13] Rojstaczer, S., & Healy, C. (2012). Where is ordinary: The evolution of American college and university grading, 1940–2009. *Teachers College Record, 114(7)*, 1-23.

[14] Schnabel, D. (2014). Patient satisfaction survey study halted; Mortality increased 238% with patient satisfaction. *Gomer Blog: Earth's Finest Medical News Site*. Retrieved 11 September 2015, from http://www.gomerblog.com/2014/08/patient-satisfaction-2/

[15] Rodabaugh, R. C. (1996). Institutional commitment to fairness in college teaching. *New Directions for Teaching and Learning, 1996(66)*, 37-45.

[16] Feldman, K. A. (1993). College students' views of male and female college teachers: Part II — evidence from students' evaluations of their classroom teachers. *Research in Higher Education, 34(2)*, 151-211.

Zhuhan Jiang received the BSc from Zhejiang University in 1982, and PhD 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 BE and ME 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.