Faculty Affairs Committee Recommendations regarding Annual Reviews and Salary Increments

The Current Practice

Several years ago, former Dean Herman Saatkamp had all the departments in the School of Liberal Arts institute a uniform annual procedure for Chairs to prepare written evaluations of their faculty members' performance. This is not the FAR that the SLA has each faculty member complete in January but rather the brief numeric scorecard for performance in research, teaching, and service that Chairs prepare for the School as a rationalization for salary increments. This score card directs Chairs to rank faculty member's performance in each of these three areas according to the following scale:

Table 1

Score	Label	Alternate label 1	Alternate label 2
4	Significantly exceeds department expectations	Excellent	A
3	Exceeds departmental expectations	Very Good	B
2	Meets department expectations	Good	C
1	Below department expectations	Satisfactory	D
0	Unsatisfactory	Unsatisfactory	F

The form still refers to the original labels, but some departments have switched to an alternate version (label 1) in an attempt to give clearer meaning to the scale. This scale is formally identical to the one we use to assign grades to our students and compute their GPA (label 2). The labels themselves are meaningful only to the extent that they provide some guideline for the assignation of a numerical score. Some Chairs have also seen fit to assign fractional scores, e.g., 2.4 (slightly exceeds Departmental expectations) or 0.8 (almost satisfactory), roughly equivalent to C+ and D- respectively. The possibility of assigning fractional scores increases the number of points on the evaluation scale almost infinitely, depending on how many decimals one is willing to consider.

The numerical scores assigned by the Chair in each category are then "weighted" according to departmental standards for the relative value of performance (weights of 0.4 each for teaching and research and 0.2 for service are common) and a final "Total Evaluation Score" is then calculated by the Chair. In addition to this final numerical rating, the form includes space for Chairs to provide brief written comments to elaborate on "whether the overall annual performance has been satisfactory or unsatisfactory with regards to both quality and productivity."

The process by which the evaluation scores are determined varies from department to department. In some cases, the Chair alone decides the scores; in other cases a small evaluation committee is charged with reviewing all the FARs and proposing scores to the

Chair; one department asks all tenured faculty to meet and take turns leaving the room to be assigned scores by everyone else. In all cases, Chairs determine the final scores, with or without input from a small or large committee.

The total evaluation score obtained is then the basis for the determination of pay increments, with Chairs allocating larger increments to faculty who received higher scores according to some formula. The following hypothetical example where 11 faculty members are allocated shares of a $22,000 envelope based on their scores illustrates how scores and pay increments are linked. The example assumes that the scores are just linear weights, i.e., $I_i = \left(\frac{s_i}{\sum_j s_j} \right) (\$22{,}000)$ where I_i is the increment received by Prof. i and s_i is his score, but other formulas are and can be used. Some Chairs have used formulas that stretch or compress the distribution of pay increments, or factor in the Faculty's current salary by allocating increments as a percentage rather than a flat amount. This can result in Faculty with equivalent performance receiving different increments if their previous salaries were unequal.

Faculty	Score	Increment
Prof. Alpha	3.8	3,074
Prof. Bravo	3.6	2,912
Prof. Charlie	3.0	2,426
Prof. Delta	2.8	2,265
Prof. Echo	2.8	2,265
Prof. Foxtrot	2.4	1,941
Prof. Golf	2.4	1,941
Prof. Hotel	2.4	1,941
Prof. India	2.2	1,779
Prof. Juliet	1.0	809
Prof. Kilo	0.8	647

Table 2

Other Chairs have simply divided the entire envelope equally among their members, and relied on Special Merit requests to the Dean to reward their top performers. It appears that Department Chairs have been making unequal use of Special Merit requests. Some have been more aggressive and/or persuasive than others in seeking and obtaining Special Merit increases for their members. This raises interdepartmental equity issues, as Deans have often awarded them unequally among Departments.

Problem with the Current Evaluation Practice

The main problem with the current evaluation practice is that it prompts Chairs and evaluation committees to make finer distinctions than available data sometimes allows, causing evaluations to be overly influenced by random elements. This can undermine morale when faculty members feel that they are being penalized or rewarded based on factors outside their control instead of their actual performance.

The following hypothetical example illustrates the concept of "insignificant statistics." Suppose that a department uses the results of a student satisfaction survey as an evaluation criterion for teaching performance. Consider the average results from three sections:

Table 3

Section	Average satisfaction (on a 5 point scale)	Department average
A	4.60	3.84
B	3.50	3.84
C	2.44	3.84

Can we conclude from this that the instructor of section A did an excellent teaching job, that the instructor of section B did an average or slightly below average job, and that the instructor of section C did a poor or unsatisfactory job? Absent other evidence, a Chair might be tempted to justify a numerical score on teaching performance based on these numbers. However, a closer look reveals that the differences in reported satisfaction may have nothing to do with the instructor's actual performance. For example, suppose all three sections were taught by the same instructor using the exact same lecture notes. Then the score differences obviously depend on factors other than the instructor's performance (unless we are willing to believe that he is so erratic that he is brilliant every Tuesdays but atrocious every Wednesdays.) To understand how, let's look at the breakdown of student answers: (in the table below the numbers in each cell show the number of students expressing each view.)

Table 4

Section	Strongly satisfied	Satisfied	Neutral	Dissatisfied	Strongly dissatisfied
A	3	2	0	0	0
B	2	2	0	0	1
C	0	3	1	2	3

We immediately notice that students who attended the same classes with the same instructor have widely different satisfaction levels. Why? Simply, because different students have different expectations, different personalities and different inclinations to express positive or negative sentiments. Since the three sections were taught in the same way by the same instructor, the logical conclusion is that the low average reported satisfaction level in section C is merely due to the random placement of students. Section C simply contained more students inclined to be critical, and fewer students inclined to be positive than section A. Knowing that the same instructor taught all three sections, a Chair could reasonably consider averaging all three scores. But what if the instructor in question had been assigned to teach only section A or only section C? No averaging would be possible, but the possibility that the high or low average student satisfaction might have nothing to do with his/her performance would remain just the same.

The question becomes more complex if the three sections were taught by different instructors, because the possibility arises then that instructor performance did affect student satisfaction. Statisticians have developed various tests to try to discern whether patterns in the data are most likely due to random variations in the sample or to a specific factor (such as instructor performance in the case here). Significance tests require that a particular pattern of data be so unlikely to be observed in a random sample (typically less than 5% of the time) that the pattern requires the presence of a significant factor to be explained. This is similar to the "reasonable doubt" test that jurors must consider when making a verdict. Here, if we cannot assert with, say, 95% confidence that the pattern of answers in all three sections could not have arisen solely because of the random assignment of students to sections, we would have to reject the hypothesis that instructor performance positively or negatively affected student satisfaction. In other words, the

data would not support the claim that the instructor of section A did a better job than the instructor of section C.

The hypothetical data presented above is an example of insignificant statistics. At first glance it seemed that the instructor of section A did a better job than the instructor of section C, but a closer look reveals that the data does not support this. The data is not "statistically significant" because it could have easily been the outcome of strictly random factors outside the instructor's control.

Similar lack of statistical significance plagues the evaluation of research performance. Many disciplines now publish journal rankings based on the average number of citations received by articles published in each journal. When prompted to make fine distinctions, the temptation is almost irresistible for Chairs and evaluation committees to jump to the conclusion that an article published in a higher ranked journal is evidence that its author had better research performance than the author of an article published in a lower ranked journal. Never mind that articles are judged not based on their own citations but based on the average citations of unrelated articles whose only connection is that they were published in the same journal; never mind that these journal rankings are based on average citations of articles published ten to five years earlier while the editorial boards of these journals have in most cases completely turned over by the time an article is accepted; never mind that the average number of citations masks the fact that different articles in the same journal typically receive widely different numbers of citations so that publication in a journal with higher average citation is no guarantee that any given article will be more widely cited; never mind that citations are only loosely linked to impact as articles get read by many people who never go on to cite them in their own work simply because their research interests and capabilities differ; and never mind that with the advent of searchable online databases researchers are just as likely to find and read an article that interests them whether it was published in a high profile or more obscure journal. Folk wisdom may hold as self evident that an article in a higher ranked journal should be worth more kudos and a bigger pay increment than one in a lower ranked journal, but in most cases hard statistical evidence just does not support this conclusion

The Faculty Affairs Committee's conclusion is that the five-point scale used for Faculty evaluation prompts Chairs and evaluation committees to make finer distinctions than can be supported by available data. This is unfair and sometimes demoralizing to Faculty who find themselves rewarded or penalized because of random events outside their control rather than because of their actual performance.

Problem with the Current Pay Practice

The main problem with the current practice of linking pay to evaluation scores is that it has perverse incentives that undermine collegiality and harm productivity. In the example in table 2, Prof. Alpha and Prof. Bravo received high evaluations. Should everyone else congratulate them? Maybe not: their good performance costs everyone else a smaller pay increment. When all pay increments come from the same pool, one's

gain is another's loss, so that good performance may be more likely to elicit jealousy than applause.

By the same token, in table 2 Prof. Juliet and Prof. Kilo have been given low evaluations and therefore received low pay increments. This benefits everyone else because every dollar not awarded to someone is freed to be awarded to someone else. Therefore others do not have incentives to help Prof. Juliet and Prof. Kilo improve their performance. In an application of the "no good deed goes unpunished" principle, the helpers' own future pay increments would be diminished if they were successful.

The committee's assessment is <u>not</u> that the School of Liberal Arts is an uncollegial place dominated by jealousies and resentments. The operative word in our assessment is "undermine." It is not healthy to have an incentive system where everyone else gains when someone is not doing a good job, because this encourages finger-pointing instead of helpful behavior. Instead of fostering collegiality and solidarity for a common purpose, the current pay practice encourages self-promotion and tearing others down.

The committee's assessment is also <u>not</u> that merit based pay is to be rejected. When properly implemented, merit pay gives everyone incentives to strive to do their best. But the current way of calculating merit pay has undesirable perverse effects. Undermining collegiality harms morale and reduces collaboration among faculty members and hence productivity. These perverse effects could be avoided by implementing merit pay differently.

Recommendation regarding evaluations

The Faculty Affairs committee recommends that the evaluation scale be simplified to three non-numerical (ordinal) categories:

- **Outstanding**
- **Meets department expectations**
- **Unsatisfactory**

We feel that most faculty members in most years meet their department's expectations but that seeking to finely classify them as slightly below or slightly above or even somewhat above would not, in most cases, be justifiable by the available data. The qualification "Outstanding" should be used to refer to a faculty member whose performance clearly stands out; someone who is so far above the crowd that this qualification is practically obvious. Similarly, we recommend that the qualification "Unsatisfactory" be applied only when there is clear incontrovertible evidence that the faculty member's performance does not meet his/her department's expectations. For example, participation in service committees is subject to random variations since there are more faculty members than committee seats in the School and campus. So non-participation by a member in any committee in any single year would not be clear evidence that this faculty member provided unsatisfactory service. However, stubborn refusal to ever contribute in any service, or service that is exceedingly spotty over several years, may represent sufficient accumulated evidence to call that faculty member's

service unsatisfactory. The committee recommends that Chairs not use the category "Unsatisfactory" to signal to a Faculty that he/she could do better, even if the faculty's performance is less than ideal. Presumably, all of us could improve something or another. For example a very good teacher could become an excellent teacher. The committee intends the category "Unsatisfactory" to be applied only to performance considered intolerably dismal. Suggestions about areas that could be improved could still be made in the written comments part of the evaluation form. Essentially, the committee intends this evaluation scale to be understood as having a fat middle and thin tails: the bar for "Outstanding" ought to be set high and the bar for "Meets Expectations" ought to be set low, so that strong and convincing evidence would be required to say that someone has either crossed the higher threshold or failed to cross the lower one.

In line with the above comment, the committee recommends that Chairs and evaluation committees try to minimize the impact of random elements by taking a long-term, rather than a strictly year-to-year, view of the faculty member's performance. For example, the vagaries of the refereeing process and publication lags mean that the time between completion of a research project and its publication can vary widely. So it would be inappropriate to say that a faculty member had unsatisfactory research performance in a year in which he published no papers, but outstanding performance the next year in which he published two. Instead, Chairs should look at the faculty's stream of accomplishments over a long period. For example, it may be that someone's performance appears strong but not outstanding in any particular year, but that viewed in total over several years, that faculty member's performance clearly stands out.

The committee recommends that Faculty continue to be evaluated separately over the three components of research, teaching and service, but that the practice of aggregating all three scores be discontinued. The scale we propose is an ordinal scale: "Outstanding" is better than "Meets department expectation" which in turn is better than "Unsatisfactory." This scale does not include an attempt to quantify the differences on a cardinal scale 4, 3, 2, 1, 0. Aggregation based on weights forces performance to be measured according to a cardinal scale. For example in the five-point scale currently used, "Excellent" is worth 33% more than "Very good" and twice as much as "Good." Under this proposal, faculty evaluations could look something like this:

Name	**Research**	**Teaching**	**Service**
Prof. Alpha	Outstanding	Meets department expectations	Meets department expectations
Prof. Bravo	Meets department expectations	Meets department expectations	Outstanding
Prof. Charlie	Meets department expectations	Meets department expectations	Meets department expectations
Prof. Delta	Meets department expectations	Meets department expectations	Meets department expectations
…			
Prof. Kilo	Meets department expectations	Unsatisfactory	Meets department expectations

An additional advantage of the simplified scale proposed here is that annual Faculty evaluations would be less time consuming. Instead of poring over the FARs to discover minute details that would justify slightly higher or lower scores, the Chairs and evaluation committees' task would be reduced to ascertaining whether anyone clearly stands out, either positively or negatively. This should be relatively easy to do as people who stand out are by definition easy to see. In most cases, evaluation committees, large or small, may no longer even be needed to inform the Chairs. The time freed from this burdensome bureaucratic procedure could then be devoted instead to improving our research, teaching and service output. This would be good for our School.

Recommendation regarding annual increments

The committee's proposal is based on the following principle: *Maintain merit based pay, while avoiding the perverse effects of the current system.* The pay scheme must also overcome the constraint that comes from not aggregating the three evaluation components: how can we then link evaluations with pay increments? The committee recommends a system that mimics the practice adopted by some Chairs of dividing general merit increments equally among their members and seeking Special Merit funds from another pot to reward their top performers. Such a practice gives everyone incentives to excel in order to receive Special Merit increments, but good performance by some does not penalize others. Special Merit could also be justified on the basis of excellence in any area. However, the flaw with this strategy is that it could not have been copied by every department. Deans have in the past allocated to departments an envelope for pay increments that they considered affordable given the School's budget, and scrounged around for some additional funds to accommodate a small number of Special Merit requests. There is no way that all outstanding performance could be rewarded this way.

However, Chairs could achieve the same effect by separating the envelope they received from the Dean in two pools of fixed size. For example, a <u>fixed</u> percentage, say, 80% or some other number to be determined by each department, of the envelope awarded for pay increments could be allocated equally among all faculty members who meet department expectations, with the rest awarded at the Chair's discretion to reward outstanding performance. What is important is that the two pools' relative sizes remain fixed: if the fraction set aside to reward outstanding performance is allowed to vary from year to year according to how many faculty members deserved it, the problem of perverse incentives would resurface. But if the two pools are fixed, then no one loses anything when someone else does well. We do not propose guidelines as to whether outstanding performance in research ought to be rewarded more, less or the same as outstanding performance in teaching or service, or whether the outstanding performance part of the envelope should be awarded to one or divided among several faculty. It is up to each Chair to award these sums according to the perceived relative merit and importance of their outstanding contributions, but clearly, Chairs should be able to justify and explain their decisions to their departments. Under this proposal, the pay increases offered to 11 faculty members sharing a $22,000 envelope could look something like this:

Table 5

Faculty	Increment
Prof. Alpha	4,600
Prof. Bravo	3,000
Prof. Charlie	1,600
Prof. Delta	1,600
Prof. Echo	1,600
Prof. Foxtrot	1,600
Prof. Golf	1,600
Prof. Hotel	1,600
Prof. India	1,600
Prof. Juliet	1,600
Prof. Kilo	1,600

Eighty percent of the envelope allocated equally to the 11 members represents $1,600 each. Professors Alpha and Bravo who were deemed outstanding in research and service respectively were awarded the other twenty percent, in this case an extra $3,000 and $1,400 respectively by their Chair. In this example, Professor Kilo's teaching was unsatisfactory but he received the same pay increment as others. The committee recommends that unsatisfactory performance not be immediately sanctioned by a smaller pay increment, but that instead this be a signal for department Chairs to set in place a corrective program. For example, why is Prof. Kilo's teaching unsatisfactory? What measures can be taken to improve it? Would assistance from the Teaching and Learning Center help? Etc… Or for a faculty whose research output is unsatisfactory, the Chair needs to identify the cause (e.g., has the faculty run out of original ideas? Is she spinning her wheels in a dead end?) and set in place corrective measures (e.g., would pairing an older faculty whose productivity has diminished with a younger faculty who is full of ideas but short of time help them both? etc…) The committee recommends that salary sanctions only be applied to chronic laggards who have resisted every attempt to help them improve their performance. Obviously, simple admonitions to "do more research" or "improve your teaching" are not what we would consider adequate corrective programs.

Comparing table 6 with table 2 above, we can also see that the committee's recommendation to separate the pay increment envelope in two strictly separate pools of fixed sizes (one divided equally and one allocated at the Chair's discretion) succeeds in eliminating the perverse incentives: No one loses when someone else does well. But then, what happens under this scheme if someone is declared a chronic laggard by his/her Chair? Then presumably the Chair could use her discretion to award that faculty member smaller or no increments from then on until performance improved. However, the perverse incentives would resurface if the freed money was then thrown back into the pot to be divvied up among other department members. To eliminate the perverse incentives, the committee recommends that the freed money be "lost" to that department, for example by returning it to the Dean who could possibly use it to set up and finance corrective programs at the School level.

What percentage of the pay envelope should be allocated to each pot? We make no recommendation about it. Each department can decide what they think is best. But there is a trade-off. Since the envelope allocated to Chairs for pay increments is often barely sufficient to compensate for inflation, then setting apart too large a portion to reward outstanding performance by some would condemn all others who are doing their job well and meet their department's expectations to real pay decreases over time. On the other hand, setting apart too small a fraction to reward outstanding performance would weaken the incentives to excel.

Naturally, any change to the grading system creates gainers and losers. The biggest gainer in this example is Prof. Alpha whose truly outstanding research performance is now more richly rewarded because the Chair is no longer constrained to reward an "excellent" score as being worth exactly 33% more than a "very good" score. The other biggest winner is Prof. Kilo who is getting a reprieve while the Chair sets in place measures to help him improve his teaching performance. Prof. Juliet is also a large gainer because the evidence is not strong enough to call her performance unsatisfactory under the simplified scale. Her low scores in the old system could have been due simply to a string of unlucky breaks. The biggest loser is Prof. Charlie whose performance did not merit an outstanding grade, but whose higher scores in the old system could have been due mostly to good luck. It is possible, however, that over a number of years, Prof. Charlie's performance could emerge as meriting an outstanding grade and be rewarded as such.

While the proposed scheme would succeed in eliminating most of the perverse incentives, it cannot eliminate them completely. The "Outstanding" portion of the pot must still be divided up among one or several meritorious faculty. Basic arithmetic constrains the amounts each could receive to be smaller if more faculty members are deemed outstanding in one area or another in any given year: It is more rewarding to be the only star in a department than to have to share this honor with several others. To lessen this problem, the committee recommends that Chairs not be constrained to fully distribute the "outstanding" portion of the increment envelope in any given year, so as to keep the budget constraint from binding. So for example, if the pot is not large enough to adequately reward outstanding performance by all those who deserve it, the chair could spread the reward over a number of years; or if in any given year, there is insufficient outstanding performance to exhaust the entire pot, the Chair should be allowed the flexibility to bank the money to be allocated later.

A final consideration is that since rewards for outstanding performance are not one-time bonuses but factored into base pay, their effect is cumulative. The committee expects that Department Chairs will continue to use good judgment in defining departmental expectations and tailoring them to individual cases so as not to let a few stars' salaries explode beyond reason. If pay is to reward performance, then more and better performance ought to be expected from faculty who are paid more. For example, a Faculty who is paid $100,000 in a department where the average salary is $50,000 could reasonably be expected to be more productive than the average, and the department could reasonably be disappointed if he produced no more than others. Similarly, a Faculty who has gone on sabbatical and was therefore relieved of teaching and service could reasonably be expected to fill his time with a commensurate increase in research activity, and the department could reasonably be disappointed if the sabbatical didn't result in a notable increase in his research output. Good judgment means that the bar ought not to be set necessarily at the same height for everyone.