1995-10

# The Evolution of a Selection System

Read, R.R.

Wiley Online

Read, R. R. "The evolution of a selection system." Naval Research Logistics (NRL)
42.7 (1995): 1099-1108.
http://hdl.handle.net/10945/59942

# The Evolution of a Selection System

**R. R. Read**
*Naval Postgraduate School, Monterey, California*

The article is a case study. It describes the initialization and subsequent modifications of the selection process used for the annual Award for Excellence in Teaching at the Naval Postgraduate School (NPS). The method treats highly unbalanced data and utilizes some exploratory data analysis techniques in interesting ways. It leads to a defensible choice for a winner in a very messy setting.
The award designates a faculty member as "teacher of the year" and includes a stipend of substantial value. The recipient is chosen by a committee which reviews objective information summarized from ballots submitted by franchised voters. The issues encountered have some general content and the handling of a number of them may have broader interest.

## 1. INTRODUCTION

This article describes the initialization and subsequent growth of a ballot processing system, specifically the selection of the recipient for the "teacher of the year" award at the Naval Postgraduate School (NPS) [6, 8]. Many of the issues involved are common to scoring systems, and the responses to them illustrate how some ideas can be implemented with data analysis [3]. The general problem is that of using judges to rate objects [1, 4, 5], and the most effective methods exploit structure and balance that are desirable but not always available. They do provide some principles, but direct application is not possible in the present case; that of a large number of heterogeneous judges rating a sizable set of diverse objects.

The award is an annual one and the judges and objects are not static; new problems arise as time progresses and sensitive modifications need to be developed in order to respond to them. The policies adopted for the governance of systems of this type depend a great deal upon the setting of the particular situation. Certain aspects of the present setting must of necessity be described, especially those that serve to direct the choice among several alternatives. The recent advances in computing capability make feasible a number of options that could not have been accomplished inexpensively in the past.

The study has three main parts. A case study must be viewed from the important aspects of the setting. Once these are understood, the initial technical aspects can be presented. Once the system is launched then the experiences gather and changes are made. This maturation process is a continual one.

Section 2 deals with the background of the particular balloting system that we have learned to deal with. Many of the issues are of general interest. Section 3 treats the initial analysis, results and experiences. Section 4 covers a rather large temporal span which in-

*Naval Research Logistics*, Vol. 42, pp. 1099–1108 (1995)

CCC 0894-069X/95/071099-10

cludes the discovery of numerous problems and the responses developed to them. A final section is included for summary purposes and some speculations about future directions.

## 2. BACKGROUND

The late sixties and early seventies was a period of considerable change as far as college level instruction was concerned. University administrators were under pressure to improve the classroom performance of instructors at all levels. The Student Instruction Report (SIR) developed by the Educational Testing Service at Princeton University came into widespread use. It was the beginning of the practice of students grading the instructors. Such feedback was intended to provide information useful to faculty so that they could improve their classroom performances.

Concurrently, an annual award for excellence in teaching was introduced at the Naval Postgraduate School (NPS) in order to augment existing incentives for the promotion of quality instruction. A committee of faculty was appointed to structure the process of selecting a winner. A timetable was set for making the first award. The award was introduced during a period of institutional crisis and it was crucial that the selection process be free from administrative bias and internal politics. Also, there was to be but a single selectee.

The committee debated for months. Ultimate conclusions included: (i) there did not exist any objective and generally accepted method for quantitatively measuring excellence in teaching; (ii) the committee could designate the recipient on the basis of the representative judgments of persons qualified to hold opinions; (iii) a polling procedure could be used to achieve this goal. Recognition was made of the fact that there were many good instructors and a variety of successful instructional styles; the committee could at best select a representative of the general level of high quality instruction. As the deadlines were drawing near, the committee proceeded to design a ballot for distribution and information processing.

NPS is a small institution of about 1600 students and over 300 faculty. The voting population needed enhancement. The committee franchised all current students, all faculty, and all alumni within two years of graduation. Eligibility rules (based upon minimal levels of instructional activity) were established to decide which faculty should appear on the ballot as candidates. Thus both the voter population and the candidates were extremely diverse.

The ballot was given a number of features. The voter was asked to supply a comparison base, i.e., to identify all those candidates with whom they were familiar as far as teaching ability was concerned. Among these the voter was asked to order his top three choices. There was also an invitation to supply a statement supporting the first choice. Such statements are not used in the selection process, but do serve to allow those responsible to learn how the franchised voting population views good teaching. The voter was also asked for some demographics, i.e., curricular area and status (student, alumni, faculty). It was decided to reject those ballots that did not identify at least five faculty candidates in their comparison base.

## 3. INITIAL ANALYSES

The ballot processing system produced four numbers for each candidate on the ballot: $x_1, x_2, x_3$, and $N$, where $x_i$ is the number of $i$ position votes received for $i = 1, 2, 3$ and $N$ is

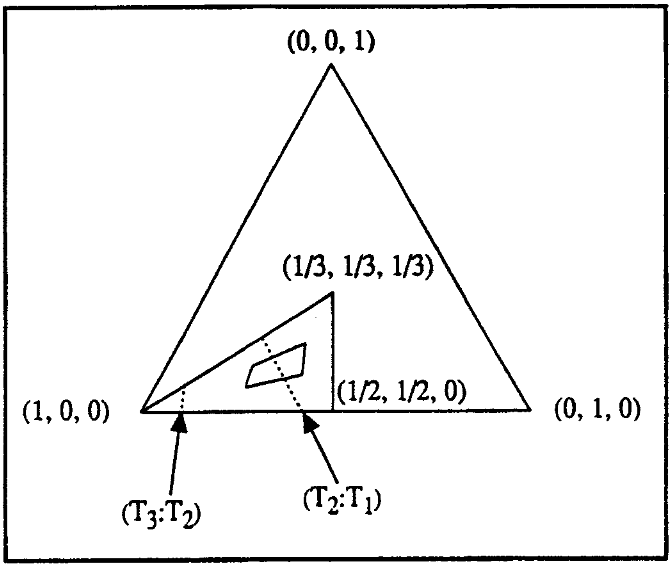

**Figure 1.** Partition of the set of admissible weights by the undominated finalists.

the number of ballots that identify the candidate in their population base. The scoring function is a weighted average of the $x$'s normalized by $N$. That is,

$$S = \sum_{1}^{3} w_i x_i / N \tag{1}$$

where $w_i \geq 0$ for all $i$.

The choice of weights was a subject of considerable debate. To resolve the debate, it was decided to treat the dual problem and to discover all those candidates who can generate a top score using any set of admissible weights. The admissible weights must satisfy the order relations

$$w_1 \geq w_2 \geq w_3 \geq 0 \tag{2}$$

with at least one strict inequality in the chain. The region is most easily described as that in the triangle subset of the simplex (without loss of generality we can require $\Sigma w_i = 1$) having corners $(1, 0, 0)$, $(\frac{1}{2}, \frac{1}{2}, 0)$, and $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. (See Figure 1). On occasion these corner points have been proposed as usable weights.

The candidates were screened using the principle of dominance borrowed from game theory. See [7] (dominance), and [2] (inadmissible strategies). A candidate with score function $S'$ was declared dominated if there exists a candidate with score function $S$ and $S \geq S'$ for all admissible weights. Then candidate $S'$ was removed from any further consideration. (Because of the separating hyperplane theorem, [2], it is necessary to check the inequality ($S \geq S'$) only at the three corners of the right triangle in Figure 1.)

The data from the first year yielded three undominated candidates. It was a relief to have so few and in subsequent years it proved to be common that there would not be many. Still the question of choosing among the three had to be faced. The nature of this condition is illustrated in Figure 1. Each of the three undominated candidates has its own subset of weight space that supports it. The three candidates are designated $T_1$, $T_2$, $T_3$ and the dashed line on the left marks those weights for which $T_2$ has the same score as $T_3$ with $T_3$ having the higher score as the weight vector moves toward (1, 0, 0). The dashed line on the right marks those weights for which $T_2$ has the same score as $T_1$ with the score of $T_1$ improving as the weight vector moves toward the other two corners of the admissible triangle. The line for which $T_1$ is tied with $T_3$ is in between these two and is of no interest. For purposes of reference, a quadrilateral of odd shape is included. Its interior marks the region in which

$$\tfrac{1}{3} \le w_2/w_1 \le \tfrac{2}{3} \qquad \text{and} \qquad \tfrac{1}{3} \le w_3/w_2 \le \tfrac{2}{3}. \tag{3}$$

The ultimate winner was selected using a paired comparison technique. Subsets of ballots were extracted in which both members of a pair of undominated candidates were identified in the population base. There were 19 ballots that identified both $T_1$ and $T_3$; 13 ballots that identified both $T_2$ and $T_3$; and one ballot that identified both $T_1$ and $T_2$. This last ballot provided no information about its originator's preference toward those two finalists, but the other two sets of ballots gave a very clear signal. The set of 13 showed a strong preference for $T_2$ over $T_3$ and the set of 19 showed a strong preference for $T_3$ over $T_1$. The committee selected $T_2$ based upon this preference chain.

Some post-award analyses were performed on the data and some of the first year results may be of interest. Positive scores were tabulated for 205 of 249 eligible faculty (82%). (This proportion increased for a few years but not much and is rather stable.) The positive scores were comfortably described with an exponential distribution having mean 0.5 (weights equal to 4:2:1 and not normalized to sum to one). Chi square tests for association of attributes did not disclose any association of the positive scores with any voter category, or voter curricular area, or the candidate's academic department.

The above approach to the selection of the award winner was continued for a few years. The dominance technique always produced a small number of finalists. The paired comparison information was helpful on occasion, but often there were either too few ballots that identified pairs of finalists in their population base, or there was not a firm signal. Always the committee deliberated in an objective manner. They studied the numerical summaries and never did they know the names of the candidates. This kind of objectivity has persisted to the present.

## 4. EXPERIENCES AND RESPONSES

A number of events in the subsequent years were disconcerting. Some faculty voters were identifying a population base of size 70 and more. This lacked credibility and, because of the nature of the scoring system, it had a negative effect upon those members of the base that were not marked in the top three. Some faculty were campaigning for votes among their students. One faculty member took remarkably extraordinary measures in his teaching style in order to win the award. He did win it and immediately reverted his teaching style to a customary one. The appearance of his name on the previous winners list lacks credibility among students that have seen his performances since then.

These and other problems appeared over a period of time. The committee's response resulted in the following changes. The faculty were disfranchised as voters. The distribution of ballots was done quietly, without publicity, so that the polling atmosphere was more dignified. An environment was promoted to make it immoral to campaign for the award. An upper limit was placed upon the size of the comparison population base allowed on a ballot. The scoring system was modified. Files of previous year scoring summaries were created and maintained. The committee used these records in their selection deliberations in order to ensure that the winner had a continuing commitment to high quality instruction. Information in the form of paired comparisons was given a formal structure. The technical details follow.

## 5. NEW SCORING SYSTEM

Under the original system, a ballot that marked a rank of $i$, $i = 1, 2, 3$, for a particular candidate contributed a value of one to the quantity $x_i$ for that candidate regardless of the size of the population base on that ballot. It seemed wise to account for the varying sizes of a ballot's population base when transferring such information to a candidate's scoring file. Accordingly, we shifted to a system of (descending) quantile scores. If a ballot identified a population base of $k$ faculty, then the first, second, and third place selections were given quantile values

$$k/(k+1); \qquad (k-1)/(k+1); \qquad (k-2)/(k+1) \tag{4}$$

respectively. In addition, all other faculty identified in the population base were awarded the average of the remaining ranks, specifically

$$(k-2)/2(k+1). \tag{5}$$

Thus the option was made available to allow a candidate's appearance in a ballot's comparison base to contribute constructively to his overall score, without that appearance necessarily being in the top three. The quantities $x_1, x_2, x_3$ were replaced by $z_1, z_2, z_3$, and $z_4$ where $z_i$ is the sum over all ballots of the quantile contribution to a faculty member's score. The new scoring formula is

$$S = \left( \sum_{i}^{4} w_i z_i \right) \Big/ N^p \tag{6}$$

where the $w_i$ are monotone weights ($w_1 \geq w_2 \geq w_3 \geq w_4 \geq 0$) with at least one strict inequality as before, and the power $p$ allows for a richer set of normalizations.

Recall that $N$ is the number of ballots that identify the candidate in their population base. It is essentially a measure of faculty exposure to students, and the scores should not be biased by it. There have been warm debates over the issue of whether or not high exposure detracts from a faculty member's ability to generate a fairly competitive score. The question can be studied using scatter plots of $S$ versus $N$. A simple regression fit produces a very flat line; the large number of mediocre and low scores dominates the fit. Figure

2 shows the scatter plots of scores for values of $p = 0.5, 0.75$, and 1.00. The data are the top 120 scores taken from the 1981 polling, and the scores are based upon the original system (three $x$'s rather than four $z$'s) with weights 4:2:1. The normalization by $N^p$ is utilized for $p = 1.0, 0.75, 0.50$. One can observe the high scores moving fluidly to the right as $p$ decreases. The visual content is that the high scores have little correlation with the exposure variable $N$ when $p = 0.75$.

It must not be forgotten that the scoring function also has an impact upon the historical records that are updated each year. These records are rather coarse in nature. At the conclusion of each annual selection process, the committee designates the top 5% of the teachers and the next 15%. The former are marked with "A" and the latter with "B." All other faculty are marked either with "E" for eligible or "I" for ineligible. The values of the exposure variable $N$ are retained also. Thus the question of there being no relationship of score with $N$ extends to the top 20% of the scores, typically from 40 to 55 faculty. We return to the treatment of the issue.

The new scoring (6) utilizes weights $w$ and the power $p$. Exploratory work has shown that they need to be chosen jointly. Recall that $w_4$ is the coefficient of $z_4$ which accumulates average quantiles for being unrated in the population base. It appears useful to allow $w_4$ to be positive. Thus there are two ways in which we counteract the negative effect of a candidate being in a ballot's population base without being ranked: There is an upper limit on the size of a ballot's base (currently 25). The averages of the ranks $1, 2, \ldots, (k-3)$ are accumulated (5) and given a positive weight.

The parameter selection process begins with the censoring of all data having too small an exposure; retaining those with $N \geq c$. This is necessary to avoid artificially inflated scores. Next we extract the values of $N$ for the top $b$ scores and the next $b$ scores. This produces two sets of exposure numbers, $N_1$ and $N_2$. The goal is to choose $w$ and $p$ so that the distributions of these two variables is essentially the same. Figure 3 carries the deciles of $N_1$ versus the deciles of $N_2$ for selected attractive parameterizations. The solid line is the identity function and is used for reference. We are seeking parameters that produce decile plots that approximate the identity function, and this is done by computer exploration.

An interpretation of a straight line fit, $N_2 = a + mN_1$, to the decile plot may be useful. The identity function has $a = 0$ and $m = 1$. This is the targeted ideal. Put $m = 1$ and $a > 0$, then the second layer of scores is for faculty whose exposures are greater than those in the first layer by about the quantity $a$. The variabilities are the same. This could support the complaint that it is more difficult to score well if you are a teacher with a larger exposure to students. The details reverse if $a < 0$. Turning to the effect of the slope while $a = 0$, a value of $m > 1$ means that the variability of the $N_2$ values is greater than that of the $N_1$ values, but the two distributions have the same location. Other parameter combinations are possible and deeper discussion of quantile plots can be found in [3]. The side parameters, $b$ and $c$, should be large enough so that the values of the scoring parameters are not particularly sensitive to their values. (For the 1993 data in Figure 3 we used $c = 10$ and $b = 25$. These appear to be satisfactory.)

The quality of fit to the identity function is measured by the root mean squared error, (rms).

$$\text{rms} = \sqrt{\sum (D_2 - D_1)^2/9} \tag{7}$$

where $D_1$ ($D_2$) are the deciles of $N_1$ ($N_2$). These values are included in Figure 3 along with the parameters. The case in the upper left has parameterization similar to that which

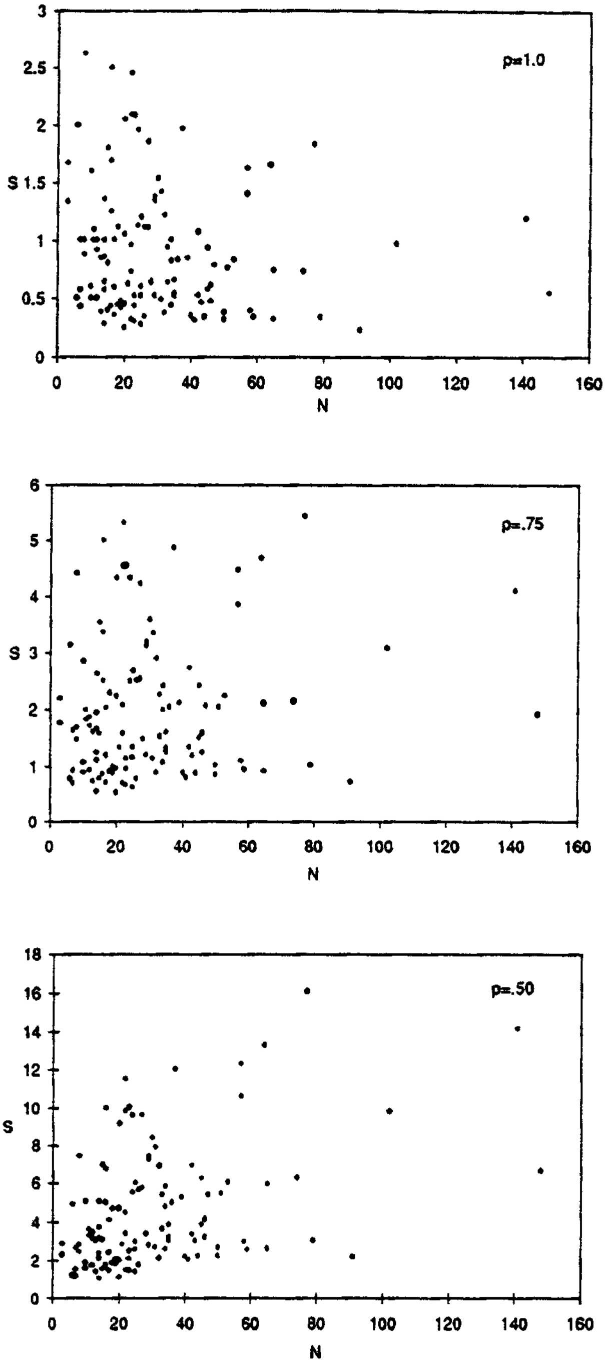

**Figure 2.** Score versus Exposure for 1981 data with $w = (4, 2, 1)$.

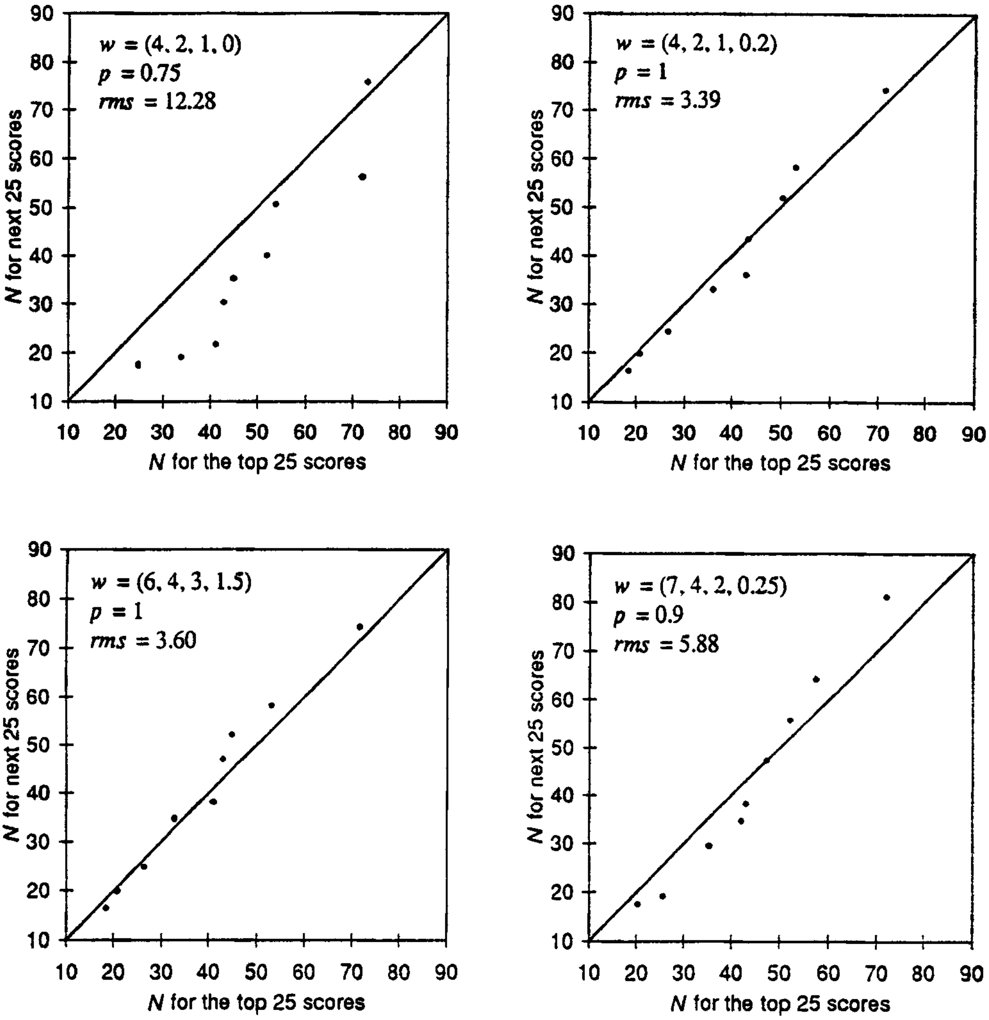

**Figure 3.** Decile plots for 1993 data.

seemed to work well in 1981, see Figure 2. But the fit in 1993, based upon the decile plot, is rather poor. The other three fits are much better as judged visually and by the rms. Indeed there are a number of parameterizations that are equally good as the one in the upper right. For most of them, $p = 1$ and $w_4 > 0$.

The year-to-year changes are noticeable and it may be desirable to make fresh parameter selection annually. But this seems to induce debate. The choice of good parameters for 1993 appears to be

$$w = (4, 2, 1, 0.2); \qquad p = 1.0 \tag{8}$$

The above considerations and the exploratory data analyses serve to illustrate the level of arbitrariness in the construction of scores from data of this type. Indeed, in picking the winner, the selection committee matches the top 15 or more scores against the past years' historical data (which has been placed in the same order as the descending scores). The earlier techniques of seeking undominated candidates and so forth were useful when there were no historical data. But now they are inappropriate.

The use of direct paired comparison information has been supplanted by the past histories. Yet a more formalized paired comparison system has been introduced and made available. The raw data for the top sixty scores are extracted and summarized into a win-loss matrix A. Its elements $a_{i,j}$ count the number of ballots on which the $i^{\text{th}}$ instructor is preferred to the $j^{\text{th}}$ instructor. Values of $\frac{1}{2}$ are added in for ties, that is, if both instructors are identified on a ballot but neither is ranked. Then these candidates are scored using the Bradley-Terry model, [2; pp. 65–66, 35–36]. The contests scored in the win-loss matrix are viewed as Bernoulli trials; a system of weights $\{ w_i \}$ is asserted to exist and possessing the property that the probability that instructor $i$ is preferred to instructor $j$ is $w_i/(w_i + w_j)$. A flat conjugate prior distribution (add 0.1 to each $a_{i,j}$ for $i \neq j$) is used; it circumvents the need to deal with a partitioning of the candidates. The weights $\{ w_i \}$ are estimated by maximizing the posterior density function. These weights serve as the paired comparison scores.

When the committee meets to perform its selection task it has three data sheets to work with: The basic contemporary scoring information $S$, $N$, and the $x$'s; the historical data as described above; and the Bradley-Terry scores placed in the proper descending order of the $S$ scores. It appears that the first two sources are used the most. The Bradley-Terry scores, although reasonably (rank) correlated with $S$, show a churning of the original rankings that generally is difficult to interpret. This, and the arbitrariness of the scoring parameters seemed to have influenced the committee to look deeper into the lists for the identification of the higher quality instructors. The paired comparison ranks have had noticeable influence in designating the A's and B's for the historical record file. The history files themselves are not used to influence the designation of A's and B's. The identification of the undominated candidates has been abandoned.

## 6. EPILOGUE

Paraphrasing a reviewer: "The paper studies a choice process that has evolved over time to select one item per year from a large number of candidates for a particular recognition. It handles all key aspects of an election based on diverse preferences and information, from the designation of eligible candidates and ballot design through scoring methodology and sensitivity analysis. Its many lessons of adaptation over time are applicable to a wide variety of choice situations that can instruct others who face similar problems."

Over the years we believe that the process has always honored high quality teachers. It has not always selected instructors who are popular with the administrators. It has always been an objective process. The popularity of the instructors with the voters is certainly a strong influence; some would say excessively so in recent times.

Other information could be added to the system now that information retrieval is far less expensive than it once was. One could devise ways to score other components such as the mechanics of teaching, the popularity of the subject area, the classroom policies of the faculty, whether the proper material was covered, and whether the students were challenged. Indeed there have been two occasions in which a committee deadlock over two top finalists was broken by a review of their grading records. All would be better if somehow we could measure the "value added" to a student for having taken a course from that instructor.

Techniques similar to those described herein might be applied in other situations that require the scoring of dissimilar objects by a large body of heterogeneous judges. The information components can be summarized objectively. The ultimate choice could be mechanized using some appropriate function or, as we have described, be reviewed by a committee whose members can debate the relative importance of the scoring components.

## REFERENCES

[1] Black, D., *The Theory of Committees and Elections,* Cambridge, New York, 1963.

[2] Berger, J.O., *Statistical Decision Theory and Bayesian Analysis,* Springer-Verlag, New York, 1985.

[3] Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P.A., *Graphical Methods for Data Analysis,* Wadsworth, Belmont, CA, 1983.

[4] David, H.A., *The Method of Paired Comparisons,* Oxford, New York, 1988.

[5] Dudewicz, E.J., and Lee, Y.J., "How to Select the Best Contender," *Annual Technical Conference Transactions of the American Society of Quality Control,* **32,** (1978).

[6] Geary, R.W., "The First Annual Award for Excellence in Teaching. A Study of Procedure and Analysis of Data," master's thesis, Naval Postgraduate School, Monterey, CA, 1970.

[7] Owen, G., *Game Theory,* Academic, New York, 1982.

[8] Read, R.R., "Data Analysis of the Teaching Award Ballots," Naval Postgraduate School technical report, NPS55-79-013, Monterey, CA, 1979.

Manuscript received August 31, 1994
Revised October 13, 1994
Accepted October 13, 1994