References and Notes

1. For a recent review, see E. Roberts and R. Hammerschlag, in *Basic Neurochemistry*, R. W. Albers, Ed. (Little, Brown, Boston, 1972), chap. 8.
2. M. Otsuka, L. L. Iverson, Z. W. Hall, E. Kravitz, *Proc. Nat. Acad. Sci. U.S.A.* **56**, 1110 (1966); L. L. Iverson, J. F. Mitchell, V. Srinivasan, *J. Physiol.* **212**, 519 (1971).
3. K. Obata, M. Ito, R. Ochi, N. Sato, *Exp. Brain Res.* **4**, 43 (1967).
4. J. J. Dreifuss, J. S. Kelly, K. Krnjevic, *ibid.* **9**, 139 (1969).
5. D. R. Curtis, A. W. Duggan, D. Felix, G. A. R. Johnston, *Nature* **226**, 1222 (1970).
6. D. R. Curtis, A. W. Duggan, G. A. R. Johnston, *Exp. Brain Res.* **12**, 547 (1971).
7. D. H. Hubel and T. N. Wiesel, *J. Physiol.* **160**, 106 (1962); H. B. Barlow, C. Blakemore, J. D. Pettigrew, *ibid.* **193**, 327 (1967); J. D. Pettigrew, T. Nikara, P. O. Bishop, *Exp. Brain Res.* **6**, 373 (1968); P. O. Bishop, J. S. Coombs, G. H. Henry, *J. Physiol.* **219**, 625 (1971).
8. D. H. Hubel and T. N. Wiesel, *J. Neurophysiol.* **28**, 229 (1965).
9. O. Creutzfeldt and M. Ito, *Exp. Brain Res.* **6**, 324 (1968).
10. R. W. Rodieck, J. D. Pettigrew, P. O. Bishop, T. Nikara, *Vision Res.* **7**, 107 (1967).
11. Bicuculline was obtained as the free base from Pierce Chemicals, St. Louis, and K & K Laboratories, Plainview, N.Y. A 1-m*M* solution was prepared as the hydrochloride form by dissolving in 0.1*M* HCl and readjusting to *p*H 6. Initial experiments gave the optimal subconvulsive dose. A dose of more than 0.3 mg/kg or very rapid injection of 0.2 mg/kg usually resulted in signs of a major convulsive episode, that is, three to four bursts or slow waves per second from the electrode and clonic muscular twitches which were visible despite paralysis.
12. K.-P. Hoffman and J. Stone, *Brain Res.* **32**, 460 (1971); J. Stone, *Invest. Ophthalmol.* **11**, 338 (1972).
13. D. Van Essen and J. Kelly, *Nature* **241**, 403 (1973).
14. L. J. Garey, *Proc. Roy. Soc. Ser. B* **179**, 21 (1971); M. Colonnier, *Brain Res.* **9**, 268 (1968).
15. D. R. Curtis and D. Felix, *Brain Res.* **34**, 301 (1971).

16. We thank H. B. Barlow for encouragement and the support of his grant EY00276-09 from the Public Health Service.

30 April 1973

Student Ratings of Teaching: Validity of Several Rating Factors

Abstract. *Students' ratings of their instructors in undergraduate classes in calculus were correlated with class performance on a common final examination. Ratings on several instructional factors were highly related to class performance even though they appeared to be independent of the students' own grades.*

Although research has rather consistently indicated a low positive correlation between students' ratings of instructors and other indices of teaching performance (*1*), the validity of student rating data remains a subject of controversy. Kossoff has questioned whether good teaching can be measured at all and has suggested that many criteria employed for students' ratings (such as friendliness, helpfulness, appearance, interest in students) may provide little information about the teacher's ability to stimulate learning (*2*). More recently, Rodin and Rodin concluded from a study of their own that "students rate most highly instructors from whom they learn least" (*3*). Because of the practical import of this issue on my own campus and because of reservations about the technical soundness of the Rodins' study, I replicated their basic research design with several modifications.

To explain my methodological concerns, a brief summary of their study will be helpful. Students enrolled in an undergraduate calculus course taught by Burton Rodin met with him in a large (293 students) lecture section 3 days a week and in small sections with one of 11 graduate teaching assistants on the other 2 days. The teaching assistants were instructed to answer questions about the lectures and homework and to administer test problems. At the end of the course, all the students took a common final examination and also rated their respective teaching assistants by responding anonymously to a list of questions about the teaching assistants' performance. Rodin and Rodin report a correlation of $-.75$ between the average rating on the item "What grade would you assign to his total teaching performance?" and the average course grade of the instructor's students. No data are presented concerning the other questions.

Although Rodin and Rodin purported to investigate their students' ability to identify good teachers, their research assessed the effectiveness of graduate teaching assistants in complementing the teaching style of one of the authors. Note also that they reported ratings on only one ill-defined global item. The correlation might have been quite different if the question had been "How well has the teaching assistant prepared you for the final examination?" They reported the mean final exam score and the mean rating for each instructor but did not report whether the differences among these sample means were reliable, that is, whether the final grades and the student ratings clearly discriminated among the instructors. If one estimates the standard error of each section's average grade from the two sections that were taught by the same teaching assistant, most of the observed differ-

Table 1. Class performance on final examination and student ratings of instructor. Ratings are on a 7-point scale, 7 being the highest.

Instructor	Mean final exam grade		Students (No.)		Instructional factor (mean rating)					
	Observed	Regressed	Taking exam	Making ratings	Work load	Student accomplishment	Organization-planning	Grading	Teacher's presentation	Teacher accessibility
					Introductory Calculus					
A	87.1	88.1	54*	38	5.52	5.02	5.64	5.42	5.10	5.10
B	84.9	84.5	38	32	5.35	4.86	4.72	5.37	4.53	4.59
C	89.2	83.9	28	23	4.20	4.96	5.28	5.48	5.16	5.74
D	85.6	83.8	34	25	4.79	5.17	4.85	5.55	5.04	6.01
E	83.4	83.3	34	27	4.27	4.75	5.22	4.97	5.35	6.33
F	77.0	79.8	24	16	5.10	4.98	4.39	5.44	4.08	5.17
H	75.6	76.9	37	26	4.63	4.63	4.46	5.33	3.56	5.63
J	70.5	72.1	46*	33	4.56	3.82	4.13	4.68	3.00	4.97
					Multidimensional Calculus					
K	77.8	76.0	19	13	4.74	4.87	5.31	5.54	4.74	6.05
M	74.0	74.1	35	27	4.21	4.75	4.91	5.02	4.37	5.52
P	71.5	72.7	48	39	3.21	4.68	5.17	5.33	5.55	5.81
S	71.2	71.4	40	32	4.22	4.41	4.63	4.58	3.81	4.72
T	62.9	62.3	37	23	3.62	4.26	4.93	4.59	3.67	5.23

* Taught in two separate sections.

Table 2. Correlation between regressed final examination score and mean rating.

Course	Instructors (No.)	Instructional factor					
		Work load	Student accomplishment	Organization planning	Grading	Teacher's presentation	Teacher accessibility
Introductory Calculus	8	.38	.84	.87	.65	.91	.14
Multidimensional Calculus	5	.51	.90	.36	.73	.60	.49
Mean		.44	.87	.62	.69	.75	.31

ences among the sections appear to be within the range one might expect from statistical sampling considerations.

In the present study two different courses were investigated, Introductory Calculus and Multidimensional Calculus. Each course was divided into classes with expected enrollments of 30 to 40 students each. Each class was taught by a regular faculty member, who met with the students for three lectures a week. There were eight instructors for Introductory Calculus (two of whom had two classes each) and five instructors for Multidimensional Calculus. The students also met once a week with a teaching assistant for a quiz section, each class being subdivided into two quiz sections. The faculty sections are the basic units of study, and the potential differential contributions of the teaching assistants are ignored.

Prior to the beginning of the fall quarter, 1972–1973, the faculty instructors in each course convened and agreed upon a common syllabus and a common text. They also met during the last week of classes to devise a common final examination. The examination papers were corrected at a common reading session by all the instructors, one instructor grading the same item for all the classes. The instructors were not aware of my research until after the last class of the quarter.

Early in the following quarter the 474 students who had completed the final exam and whose Scholastic Aptitude Test (SAT) scores were on file at the university were sent an instructional rating form by mail. Completed rating forms were returned by 354 students. The form consisted of 18 statements to be scored on a 7-point scale. It had been developed as a general instrument for the entire university, and therefore the statements did not relate specifically to calculus classes. The returned forms from these and other classes ($n = 575$) were factor-analyzed with the set of responses from each student as the basis of the inter-item correlation matrix. This analysis involved a principal-components solution followed by varimax rotation (*4*) and clearly indicated that the 18 items could be grouped into six factors, each composed of three items. These factors have been designated as student accomplishment (sample item: "this course has developed my ability to examine the evidence in this field"), work load ("this course had a heavy work load"), organization-planning ("the details of this course were carefully planned in advance"), grading ("the grading procedure in this course was fair and impartial"), teacher's presentation ("the teacher communicated his ideas in an unambiguous manner"), and teacher accessibility ("the teacher listened to students' questions and was willing to help"). On the basis of the factor analysis I assumed that the questionnaire provided information about six aspects (dimensions) of the classroom experience, and therefore the rating data for each factor are averaged across the three items which compose it.

A regressed final examination score was calculated for each student by taking the difference between his observed score and the score predicted on the basis of his SAT profile and adding this to the average observed score of all the students taking that examination. The average regressed final exam score for each instructor's class (Table 1) is used as the external criterion for testing the validity of the instructional ratings. The reliability of the differences among instructors was assessed by one-way analyses of variance. The instructors were a significant source of variance on their students' regressed examination scores in both Introductory Calculus ($F = 5.51$; d.f. = 7, 287; $P < .001$) and Multidimensional Calculus ($F = 3.57$; d.f. = 4, 174; $P < .01$). Table 1 also presents the average ratings for each of the 13 instructors on the six rating factors. The instructors were a significant source of variance on all six factors, with F ratios of 13.16, 10.63, 8.21, 5.26, 4.05, and 3.25 for teacher's presentation, work load, teacher accessibility, organization-planning, student accomplishment, and grading, respectively (d.f. = 12, 341, $P < .001$ in each case). This result indicates reasonably high interrater agreement on the teacher's presentation and work load factors.

Pearson product-moment correlation coefficients were calculated between the average factor ratings for each instructor and the average regressed final exam score of his students. The observed correlations (Table 2) were positive in all cases. The student accomplishment factor correlated significantly in both Introductory Calculus ($r = .84$, d.f. = 6, $P < .01$) and Multidimensional Calculus ($r = .90$, d.f. = 3, $P < .05$). Since correlation coefficients based on such a small number of observations are notoriously unstable, the mean correlation for the two courses (Table 2) is probably the best estimate of the strength of association between each rating factor and final exam performance. In those data, in addition to the student accomplishment factor, the teacher's presentation factor also correlated highly. Work load and teacher accessibility were not as useful in predicting exam performance.

The student accomplishment factor provides a students' estimate of how much was learned in the course; therefore it seems reasonable that it should correlate highly with a more objective measure of how much was learned. In fact, this relationship would seem almost trivial if it were not for the seemingly contradictory evidence reported by Rodin and Rodin. The relationships I find most interesting are those between the ratings of specific aspects of teaching and final exam scores. For example, it would appear that good teaching in calculus consists of presenting ideas and concepts clearly (teacher's presentation) and of planning class time well (organization-planning). The personal attention available to students (teacher accessibility) and the amount of work required of them (work load) seem to be of lesser importance. However, one can be misled by only considering these factors individually. For example, the work load and teacher's presentation ratings taken together correlate .98 with exam scores in Introductory Calculus and .92 with exam scores in Multidimensional Calculus. Since the simple combination of the two factors is a better predictor than either one alone, it would appear that there is a trade-off relationship between the clarity of an instructor's

presentations and the work load he imposes on his students; a shortage of one can be compensated by an increase of the other.

Because the ratings were obtained after the students had received their final grades, it might be argued that the statistical associations are an artifact of the students' reactions to their grades, such as a desire to "get even" with instructors who give them poor grades. This "retaliation hypothesis" can be tested in the data in two different ways:

Students who received low grades might have been more likely to return the mail questionnaire and also more likely to give their instructor low ratings. But in fact, the average nonregressed exam grade for the 222 students in Introductory Calculus who returned the rating form was 84.2 and for the 81 who did not was 72.6. This difference is opposite to that predicted by the "retaliation hypothesis" and is highly reliable ($t = 5.16$, d.f. $= 301$, $P < .001$). Similar data were observed in Multidimensional Calculus: an average grade of 73.8 for the 132 students who responded and 63.3 for the 56 nonresponders ($t = 3.93$, d.f. $= 186$, $P < .001$).

Second, if grades are causally related to ratings independent of the teacher's performance, there should be a high positive correlation within each class between a student's final exam grade and his ratings of the instructor. In the present study these correlations have been calculated for the two factors showing the highest interrater agreement, work load and teacher's presentation. They range from $-.33$ to $+.43$ with an average value of $-.02$; five are positive and eight are negative. These outcomes are not consistent with the hypothesis in question; there is no evidence for a strong positive relationship between final exam grades and the ratings when the effect of the different instructors is removed.

A reasonable explanation for the differences between my results and those of Rodin and Rodin can be formulated by considering the differences in our methodologies. The negative relationship they observed may be a unique outcome which was highly dependent on the principal lecturer's teaching style and the way this style affected the performance of his teaching assistants. Second, the Rodins' rating measure required the students to make a global judgment about teaching performance whereas my questions focused on more discrete aspects of teaching and on observable behaviors. I believe that the very strong relationships in my study resulted from a successful effort to categorize student ratings in terms of specific factors and thus to be able to separate more useful from less useful ratings. Further research with separate factors might make it possible to identify the important aspects of teaching in particular fields.

PETER W. FREY
Cresap Neuroscience Laboratory, Northwestern University, Evanston, Illinois 60201

References and Notes

1. H. H. Remmers, F. D. Martin, D. N. Elliot, *Purdue Univ. Stud. Higher Educ.* **66**, 17 (1949); D. N. Elliot, *ibid.* **70**, 5 (1950); R. J. Wherry, *PRB Report 921* (Personnel Research Branch, Personnel Research and Procedures Division, Adjutant General's Office, Department of the Army, Washington, D.C., 1952); M. T. Miller, *J. Educ. Psychol.* **62**, 3 (1971); J. E. Morsh, G. G. Burgess, P. N. Smith, *ibid.* **47**, 79 (1956); P. K. Gessner, *Science* **180**, 566 (1973).
2. E. Kossoff, *Amer. Scholar* **41**, 79 (1971).
3. M. Rodin and B. Rodin, *Science* **177**, 1164 (1972).
4. A sample rating form and a copy of the factor matrix may be obtained from the author.
5. Supported by the Center for the Teaching Professions, Northwestern University, which is funded by the Kellogg Foundation. I thank B. Claude Mathis, director of the center, for helpful suggestions and encouragement.

25 June 1973; revised 9 August 1973

Origin of Mitochondria

In relation to the recent comments between Uzzell and Spolsky (*1*) and Raff and Mahler (*2*), I wish to propose (*3*) that the primitive phagocyte in which bacterial ancestors of mitochondria allegedly settled some 1.5 billion years ago was actually an aerobic cell that relied on peroxisomes instead of on a phosphorylating electron transport chain for its respiratory metabolism. This hypothesis was formulated mainly in an effort to retrace the evolutionary history of the peroxisome, a particle which certain facts suggest may have been of much greater metabolic importance in early eukaryotes than it is in many plant and animal cells today. Acquisition of the more efficient mitochondria was put forward as an explanation of the evolutionary decline of the peroxisome. By the same token, possession of a primitive respiratory system would have made acquisition of mitochondria advantageous even in an aerobic cell. Thus the objection that "the aerobic nature of the ancestral protoeukaryotic cell would make the acquisition of an aerobic symbiont unnecessary" (*2*) loses much of its pertinence.

According to my hypothesis, both the endosymbiont and its host are pictured as originating from primitive aerobic bacteria endowed with a peroxisomal type of respiration ($O_2 \rightarrow H_2O_2 \rightarrow 2\ H_2O$), one evolutionary line leading to the development of a respiratory chain and of coupled phosphorylating systems, the other to the acquisition of phagocytosis and intracellular digestion, proliferation of intracellular membranes, and an increase in cell size. It will be noted that these two evolutionary lines correspond to two distinct, and possibly mutually incompatible, differentiations of the cell membrane.

C. DE DUVE
Laboratoire de Chimie Physiologique, Université de Louvain, 3000–Louvain, Belgium, and *Rockefeller University, York Avenue and East 66 Street, New York 10021*

References

1. T. Uzzell and C. Spolsky, *Science* **180**, 516 (1973).
2. R. A. Raff and H. R. Mahler, *ibid.* **177**, 575 (1972); *ibid.* **180**, 517 (1973).
3. C. de Duve, *Ann. N.Y. Acad. Sci.* **168**, 369 (1969).

11 May 1973

Internal Gravity Wave–Mean Wind Interaction

Bekofske and Liu (*1*) have demonstrated that the interaction of a vertically propagating internal gravity wave (IGW) with the background wind shear near a critical level (where the mean wind speed equals the phase speed of the IGW) can increase the background wind shear sufficiently to satisfy the criterion for Kelvin-Helmholtz instability. Breaking of the resulting Kelvin-Helmholtz waves would then be expected to produce clear air turbulence (CAT). This mechanism is indeed a plausible source for some CAT. However, the idea is not a new one (*2*).

Numerical calculations quite similar to those in (*1*) have previously been reported by Lindzen and Holton (*3*) in their study of the quasi-biennial oscillation in the mean zonal wind in the equatorial stratosphere. Lindzen