attended by tachycardia (90 to 160 beats per minute); this was followed by a breath hold for 20 seconds to 4 minutes, during which the heart rate slowed to 30 to 70 beats per minute. In some animals the transition between REM and slow wave sleep (SWS) was abrupt, taking 1 minute or less; in others, REM and SWS alternated (along with heart and breathing rates) for up to 30 minutes before SWS became dominant and sustained.

Slow wave sleep that lasted for 20 minutes to 4 hours was observed whether the seal was completely out of the water, in the water with just the tip of the nose clear during breathing, in the water bobbing up and down, or in the water with the head out. The observer could approach the seal and speak quietly without inducing behavioral arousal. If the observer touched the seal, of if a seal in another tank nearby vocalized, behavioral arousal followed.

During the 12 all-night sessions an average of 2.8 ± 1.5 hours was spent in active waking (AW), 5.0 ± 2.3 hours in QW, 4.7 ± 2.2 hours in SWS, and 1.5 ± 1.1 hours in REM (*12*). When seals awoke spontaneously and became active during the night, a return to behavioral quiescence always followed the sequence QW to REM to SWS. When seals awoke momentarily from SWS they returned to that state.

The sleep of gray seals was distinctive from that of man and other terrestrial mammals that have been similarly studied in that (i) it took place in water as well as out; (ii) REM sleep was accompanied by a rapid, regular heart rate, whereas in other mammals studied it is accompanied by the lowest and most irregular heart rate (*13*); (iii) respiration was regular during REM sleep, whereas in terrestrial mammals it is irregular; and (iv) REM sleep in the seal appeared first, whereas in other mammals it normally follows SWS. The gray seal, and perhaps other pinnipeds as well, may have evolved a unique sleep mechanism to cope with the necessity for sleeping in water.

S. H. RIDGWAY
Biosystems Research Department, Naval Undersea Center, San Diego, California 92132
R. J. HARRISON
Department of Anatomy, University of Cambridge, Cambridge, England
P. L. JOYCE
Department of Physiology, University of Cambridge

References and Notes

1. M. C. Caldwell and D. K. Caldwell, in *Mammals of the Sea*, S. H. Ridgway, Ed. (Thomas, Springfield, Ill., 1972), p. 450.
2. M. Jouvet, *Physiol. Rev.* **47**, 117 (1967).
3. R. S. Peterson and G. A. Bartholomew, *The Natural History and Behavior of the California Sea Lion* (American Society of Mammalogists, Oklahoma State Univ., Stillwater, 1967).
4. K. M. Backhouse, *Seals* (Baker, London, 1969), p. 23.
5. H. Hediger, *Studies on Psychology and Behavior of Captive Animals in Zoos and Circuses* (Criterion, New York, 1955), p. 25.
6. S. H. Ridgway, in *Mammals of the Sea*, S. H. Ridgway, Ed. (Thomas, Springfield, Ill., 1972), p. 599; Y. C. Lin, D. T. Matsuura, G. C. Whittow, *Am. J. Physiol.* **222**, 260 (1972).
7. R. W. Pierce, thesis, University of California, Berkeley (1970).
8. T. B. Fryer and G. J. Deboo, in *Engineering in Medicine and Biology* (Institute of Electrical and Electronics Engineers, New York, 1964); T. B. Fryer, H. Sandler, B. Datnow, in *Proceedings of the 7th International Conference on Medicine and Biology in Engineering* (Stockholm, 1967); *Med. Res. Eng.* **8**, 9 (1969).
9. Six gray seals obtained in Iceland were employed in these experiments. Electrophysiological observations were made on four of the seals. The seals were acquired in November 1970, shortly after they were weaned and had begun taking dead fish in captivity. The experiments were conducted during late 1971 and early 1972, when the seals weighed 70 to 100 kg and were well adapted to captivity. All experiments were conducted at the University of Cambridge under a license from the British Home Office.
10. Since FM telemetry is virtually useless in sea water, we kept the seals in fresh water. Gray seals have been maintained for long periods of time in fresh water in various zoos. Each seal was kept in an individual tank to which salt was added for a few days after each implant.
11. A portable FM radio receiver (Sony, model 7F-74DL) was placed within 5 m of the implanted seal. The output of the radio was connected by an electrical lead to a decoder and displayed on an oscilloscope (Tektronix, model 564B) and a polygraph (Grass, model 78). The transmitters and the decoders were constructed in our laboratory.
12. Averages were rounded off to the nearest tenth of an hour.
13. W. Baust and B. Bohnert, *Exp. Brain Res.* **7**, 169 (1969); W. Baust, J. Bohmke, U. Blossfeld, *ibid.* **12**, 370 (1971).
14. We thank Derek Thurlborn and Andrew Greenwood for technical assistance and William F. Flanigan for reviewing the manuscript.

2 July 1974

Student Evaluation

Gessner (*1*) and Frey (*2*) appear to contradict our finding (*3*) of a substantial inverse correlation between amount learned by students and evaluation of the instructor. However, methodological and substantive aspects of both their studies undercut the superficial contradiction.

Gessner reports a positive correlation between rating of instructor and amount learned and concludes that student ratings are a good measure of teaching performance. However, instructor and subject matter are confounded in the Gessner study. Assume for a moment an acceptable measure of amount learned. The different instructors taught different subject matters. One could just as reasonably conclude from the positive correlation between instructor rating and amount learned that students perform better in subject matters they like better, quite independently of who is teaching or how.

The preceding assumption of an acceptable measure of amount learned is not, however, warranted. Gessner's assertion that little could be concluded from student performance on departmental examinations is clearly correct. Since the various instructors taught different subject matters and each prepared his own questions, it would have been impossible to tell whether students did well in a subject area because the instructor's teaching was superior or because his questions were easier. Gessner rejected student performance on departmental exams as a measure of student learning for rather different reasons: they did not correlate positively with student ratings and they were uncorrelated with student performance on a national exam (the National Medical Board Examination). He therefore used performance on the national exam as his measure of amount learned. More precisely, he measured student learning by how well his sample of students performed relative to the national sample. Gessner's use of this measure presents several problems.

Gessner takes the performance of the national sample on a given question as a measure of the difficulty of the question. However, he found no relationship between performance of the national sample and performance of his sample in the individual questions of the national exam [reference 18 in (*1*)]. The finding that his measure of item difficulty is unrelated to actual item difficulty for his sample would seem to invalidate his measure. Performance on the national exam is a good measure of item difficulty only on the assumption of standardized course content. Otherwise, unusually high or low scores obtained by a student sample may simply reflect unusual content emphasis relative to the norm.

The most serious problem with Gessner's measure of student learning is that it takes no account of what the instructor fails to teach. Assume that each subject matter unit is allotted an average of two lectures. Consider the following hypothetical situation. Instructor A chooses to spend his two lecture hours on a single point. Instructor B is more ambitious and decides to spend his two lecture hours covering ten different points. The national exam has questions on all eleven points. Assume the performance of the national sample is 50 percent. One-hundred percent of the students in the experimental sample pass the question on A's material. Fifty percent of the students in the experimental sample pass each of the questions on B's material; this outcome results from the fact that each student learned five of the ten points perfectly and failed to learn the other five. Using Gessner's measure of performance of the experimental sample relative to the national sample, instructor A obtains a score of +50, and instructor B a score of 0, although the students learned five times as much from instructor B. In short, Gessner's measure makes it possible for the least amount learned to look like the most amount learned.

Gessner raises several points about our study. He says that our measure of student learning did not take the rate of learning into account and hence the opinions of slower students were given undue weight. He is quite correct that the measure was of total amount learned, independent of the rate at which it was learned. However, we see this as a virtue rather than a defect. Rate of learning is the poorer measure because it is more closely tied to the factor of student ability: within a highly selected college population, although the brighter students learn faster, they do not necessarily learn more. Second, it is not at all clear that the opinions of the slower learning students about their instructors should carry less weight than those of the faster students; surely their learning is no less important.

Gessner also notes that "if no significant correlation is found between student ratings and class performance on examinations, then there would appear to be no a priori basis for singling out, as Rodin and Rodin . . . have done, one of the variables as reflecting teaching effectiveness more accurately." In fact, a significant correlation (albeit negative) was found. Aside from that, however, he makes an important point —one that deserves a more explicit elaboration. We defined two criteria of teaching effectiveness: a subjective criterion of student ratings and an objective criterion of amount learned. These terms were intended to be merely descriptive: there was no implication of differential value. If amount learned and student ratings give two different orderings of instructors, then it is pointless to ask which ordering is valid. The judgment of which ordering is most useful will depend on educational values and purposes extraneous to the ordering itself.

It is worth noting the unusual logic of Gessner's concluding paragraph. He notes that there is a positive relationship between student ratings and class performance on national exams, but no relationship between student ratings and class performance on institutional exams. He concludes from these findings that student ratings and class performance on national exams are valid measures of teaching effectiveness. His conclusion is unwarranted, even if it is not the non sequitur it appears to be. The confounding of instructor and subject matter, and the difficulties with the measure of what students learned, cannot be ignored.

Frey claims to have constructed a teacher evaluation scale that measures independent aspects of teaching performance and yields a high positive correlation between student learning and teacher evaluations (that is not a statistical artifact). Both these claims are questionable.

Frey criticizes our study because the student ratings were obtained "on only one ill-defined global item." Our decision to use a global rating, as was explained in our article, stemmed from an examination of the work of Remmers and his associates (*4, 5*) using the carefully developed Purdue Rating Scale. Attempts to find those items on the scale that discriminated good from poor instructors (as defined by student performance) indicated that only the global rating item consistently differentiated between these two groups of instructors.

I am in sympathy with the point that it makes more sense to obtain differentiated ratings. Frey, however, has not succeeded in obtaining them. Table 2 of his report presents the six factors which presumably underlie student responses. If the factors are independent, as is the implicit claim, they could together account for no more than 100 percent of the variance in final examination scores—each factor would account for a unique component of the criterion. This is clearly not the case. If mean ratings on the separate factors are combined additively to predict final examination scores, they account for 286 percent of the variance. The table makes sense only if the six factors all load heavily on a general underlying factor. Frey has not resolved the global rating problem. His factors are not independent—they are highly intercorrelated.

Frey collected the teacher evaluations some weeks ("early in the following quarter") after the students had received their grades. As Frey points out, it is not at all surprising to find a high correlation between student ratings of their accomplishment and their grades when they rate their accomplishment after having been informed of their grades. It is equally unsurprising to find that student accomplishment also correlates highly with the other five factors. This is inevitable, since the factors are not independent. The high positive correlations he obtained between his factors and student accomplishment may, in other words, reflect no more than the timing.

In discussing the timing problem, Frey states, "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." He terms this the "retaliation hypothesis." Frey does not mention the logical complement to the "retaliation" hypothesis, which can be called the "reward" hypothesis—the desire of students to reward instructors who give them good grades. Frey counters the retaliation hypothesis by noting that the mean grades of those students who responded were not worse than the mean grades of their respective sections, but were, in fact, significantly better. Although this finding is convincing evidence against the "retaliation" hypothesis, it is equally convincing evidence for the "reward" hypothesis. In short, the finding that responders' grades differed significantly

from those of the total student sample makes it seem very likely that the statistical associations were an artifact of the students' reaction to their grades.

His second argument turns on the point that it cannot be argued that ratings simply reflect grades because there was not a consistent within-class correlation between grades and ratings (they ranged from $-.33$ to $+.43$). However, within-class data is irrelevant to the general point. The across-class and within-class correlations are independent: it is quite possible for the correlation across classes to be positive (or negative) and for the correlations within classes to be positive, negative, or both. Our study suggested that those instructors whose students learned least, on the average, tended to get higher average ratings. This finding implies nothing about whether the poorer or better students will like any given instructor better. The suggestion of Remmers *et al.* (*4*), that it depends on the level at which he pitches his teaching, seems to be a very plausible one.

It may well be the case that high positive correlations sometimes obtain between teacher ratings and amount learned, although Gessner's article and Frey's report do not appear to adequately demonstrate that point. If high positive correlations sometimes obtain, this in no way negates the fact that high negative correlations sometimes obtain. The point is no longer to demonstrate that both can happen, but to understand the circumstances under which each will happen.

MIRIAM RODIN

Department of Psychology,
San Diego State University,
San Diego, California 92115

References and Notes

1. P. K. Gessner, *Science* **180**, 566 (1973).
2. P. W. Frey, *ibid.* **182**, 83 (1973).
3. M. Rodin and B. Rodin, *ibid.* **177**, 1164 (1972).
4. H. H. Remmers, F. D. Martin, D. N. Elliot, *Purdue Univ. Stud. Higher Educ.* **66**, 17 (1949).
5. D. N. Elliot, *ibid.* **70**, 5 (1950).
6. My thanks to Rebecca Bryson for her very helpful comments.

5 November 1973; revised 22 April 1974

Miriam Rodin ends her comment by suggesting that both positive and negative correlations can be observed between student ratings of instruction and an external performance criterion. On this point we are in complete agreement. Neither my study nor Gessner's study contradicts her findings. Although my research is clearly more similar to the Rodins' than is Gessner's, both of our studies involve major methodological changes, and therefore the different outcomes require a careful consideration of the differing methodologies. As Rodin concludes, "The point is no longer to demonstrate that both [positive and negative correlations] can happen, but to understand the circumstances under which each will happen." The major point of my report was to suggest that the Rodins' surprising findings were understandable given the circumstances of their study.

My report mentioned three major differences between the Rodins' study and mine. One concerned the operational definition of teaching. The Rodins' "teachers" were graduate students whose responsibilities consisted of giving quizzes and drilling the students on calculus problems. Burton Rodin was responsible for the organization of the course and for the three main lectures delivered each week. Good teaching in this context would seem to reflect how well each graduate student complemented Burton Rodin's teaching style and how well he motivated the students to meet Rodin's goals. In contrast, my study involved regular faculty members who organized and conducted their classes in a manner which satisfied their own preferences.

Second, as Miriam Rodin mentions, my questionnaire elicited information about several different aspects of each course while her study involved a single global rating item. Although she is "in sympathy" with my attempt "to obtain differentiated ratings," she believes the attempt was a failure. My several rating factors do contain considerable common variance, as she points out, but they also each involve a substantial unique component. Factor analyses on several sets of data from different disciplines and from different universities have indicated that my questionnaire has a robust, easily replicated factor structure. The important point in this matter, however, is obscured by Rodin's emphasis on the differences between a multiple-item and single-item questionnaire. If the Rodins had selected a reasonable single item, I would have been less critical of their dependent measure in my report. The fact of the matter is, however, that an item such as "what grade would you assign to his total teaching performance" bears very little relationship to the Rodins' external performance criterion. A better choice might have been an item such as "how well has the teaching assistant prepared you for the final exam." I would be surprised if this latter item led to a negative correlation. The factor in my study which correlated most highly with the external criterion was student accomplishment ($r = .87$). This factor incorporates items such as "this course has increased my knowledge and competence in this area" and " this course has developed my ability to analyze issues in this field." In subsequent validity research with this questionnaire involving multiple-course sections using a common course outline, a common text, and a common final exam, the ratings on the student accomplishment factor have consistently correlated well with the external performance criterion [12 sections of calculus at Northwestern University ($r = .61$); 9 sections of educational psychology at Purdue University ($r = .53$); 5 sections of calculus at North Dakota State University ($r = .64$)].

A third major difference in our two studies was the time when the ratings were collected. My study was unusual in that the ratings were collected by a mail survey approximately 6 weeks after the end of the course. In most studies, including the Rodins' study and my subsequent studies mentioned above, the ratings have been collected during the last week of the term. Rodin suggests that my high correlation simply reflects the fact that students receiving good grades gave their instructors high ratings and students receiving poor grades gave low ratings. This is a plausible hypothesis, as the students' final grades would be positively correlated with final exam scores. Because I was aware of this possibility when I conducted the study, I examined data which explicitly tested this hypothesis. Since the correlation between final exam scores and instructor ratings calculated between sections is completely independent of this same relationship calculated within each section (as Rodin points out), it is possible to determine if the ratings depend upon the instructor's teaching performance or upon the grade which was given to the student. If my ratings were simply a reflection of the grades received, then this relationship should be observed within each section. The within-section correlations are tests of this relationship with the differences among teachers removed as a source of vari-

ance. The between-section correlation includes the teacher differences as a source of variance. The strong between-section correlation in the absence of any consistent within-section correlations completely refutes Rodin's contention.

Although Rodin believes that my timing was an unfortunate design error, I am presently leaning toward the viewpoint that it might have been a very fortuitous decision. The common procedure of collecting instructor ratings in the classroom at the end of the term is more a matter of convenience than a choice based on considerations concerning the reliability or validity of the data. As previously mentioned, my basic research design has subsequently been repeated three times with the exception that the ratings were collected during the last week of the term in each case. Although the correlations between the ratings and the external criterion were consistently positive, the strength of the relationship was not as great as that observed in the study I reported in *Science*. Based on these observations, it seems plausible to hypothesize that ratings collected a month or so after the course has ended may be intrinsically more valid.

PETER W. FREY
Department of Psychology,
Northwestern University,
Evanston, Illinois 60201
19 July 1974

Rodin's response fails to address my main criticism of the Rodins' study (*1*). They claimed that their evaluative procedure reflected how much the students learned from the instructors they were rating. Yet the instructors rated were teaching assistants who gave no lectures, while the students spent 60 percent of the course time being instructed in lecture by a professor. Moreover, the teaching assistants' role was limited to answering questions about the professor's lecture, going over homework assigned by the professor, and administering paradigm problems devised by the professor. Thus, the teaching assistents had little scope for developing their own teaching strategies. Therefore, presumably, the mean grade achieved by their section reflected, at best, on the ability of the teaching assistants to coordinate and dovetail their efforts with those of the professor. Accordingly, Rodin and Rodin's results reflect not, as they contend, a simple interaction between students and teaching assistants, but rather a three-sided one between students, teaching assistants, and the professor. Because of this, the negative correlation reported by the Rodins may have been due to factors other than those discussed by them. For instance, those students who found the professor's teaching approach least effective may have tended to rate highest those teaching assistants who departed most from the professors approach; yet such students could be expected not to do as well on the evaluative device (the paradigm problems) set up by the professor. In my discussion (*2*) of Rodin and Rodin's article, I suggested that details which "could shed further light on this would be of interest. How, for instance," I asked, "did the ratings of the assistants and the professor by the students in the various recitation sections compare?" Rodin does not provide this information, nor does she deal with this criticism.

There is another problem with Rodin and Rodin's study. In addition to teaching assistants and students in their recitation sections being confounded, the students were not assigned to the various recitation sections at random but could choose sections and teaching assistants as they liked. The process was made even less random by the students having had prior contact with the teaching assistants. Accordingly, students who did least well in the previous quarter, and had thereby most cause for dissatisfaction with their teaching assistant, would be the ones most likely to change sections. It could be predicted that such students, as a whole, would also not do so well in the following quarter. Yet when asked to rate their newly chosen teaching assistant, they were being asked, in effect, to also pass judgment on their own act of choosing. Such personal involvement would likely bias the ratings these students might make of their teaching assistants. Rodin and Rodin, in an attempt to correct for the confounding and for any bias introduced by the nonrandom selection procedure, calculated for each section an index of initial ability based on the previous quarter's grade. It is quite doubtful that this index, however, could correct for the additional bias discussed above.

Rodin's primary criticism of my study is based on an apparent misreading by her of reference 18 in my article. The purpose of my study was to determine whether there was substantial positive correlation between the ratings students gave to instruction in various subject areas and class performance in these areas on examinations. Instruction in all subject areas was rated by the same group of students. Class performance, however, was necessarily determined on the basis of different questions in each subject area. Since the intrinsic difficulty of the questions varied, it was necessary to allow for this. I used the percentage of the nationwide sample answering any given question correctly as a measure of question difficulty. Specifically, class performance on the national examination in any given subject area j was calculated as the average difference between the score my sample and the nationwide sample obtained on questions in that area, that is, as

$$P_{\cdot j} = \frac{\sum_{i=1}^{n}(y_{ij} - z_{ij})}{n} \quad (1)$$

where y_{ij} and z_{ij} are the percentages of the class and the nationwide sample, respectively, answering question i in subject area j correctly, and n is the number of questions in subject area j.

Rodin suggests (third and fourth paragraph of her comment) that my use of the percentage of the nationwide sample answering a question correctly (that is, of the quantity z_{ij}) as a measure of question difficulty is invalid. She apparently interprets my reference 18 as reporting a complete lack of correlation between the percentage of the nationwide sample and my sample answering each of the 141 questions correctly, or specifically

$$r_{(z_{ij},\, y_{ij})} \quad (2)$$

Careful reading of my reference 18 shows, however, that the correlation coefficient reported therein is not correlation coefficient 2 above but rather

$$r_{(z_{\cdot j},\, P_{\cdot j})} \quad (3)$$

that is, one between class performance ($P_{\cdot j}$) as defined in Eq. 1 and the mean percentage of the nationwide sample answering the questions in subject area j correctly, which was computed as

$$z_{\cdot j} = \frac{\sum_{i=1}^{n} z_{ij}}{n} \quad (4)$$

and that, therefore, the value of correlation coefficient 3 is not germane to the criticism Rodin advances. Moreover, I have now computed, using the original

raw data upon which my study was based, the correlation coefficient 2 which is germane to her criticisms. Far from being equal to zero, this had a value of .86 and is highly significant ($P < .001$). Had its value been zero, as suggested by Rodin, that would have implied that my student or instructor sample, or both, were totally unrepresentative of their respective nationwide populations.

Another criticism Rodin advances (fifth paragraph of her comment) is that some of the questions on the national examination were excluded from my study because the subject matter was not covered in the course. However, as stated in reference 15 of my article, only a relatively small proportion (8.3 percent) of the questions were thus excluded. Second, with the possible exception of two borderline cases representing only 1.4 percent of the total, questions were not excluded for the reason that an instructor omitted the relevant material. Rather the rationale for excluding questions was that they pertained to certain special subject areas in which the department had decided not to offer any instruction at all and had so advised the students. Had questions been excluded from the various subject areas of my study in the manner suggested by Rodin's example, the partial correlation coefficient for the student ratings and class performance, with relative emphasis held constant, would have been markedly smaller than reported in my article. Specifically, exclusion of questions in this manner would have had a marked effect on the relative emphasis accorded to that subject area, this being calculated in my study as

$$E_j = \frac{a_j - b_j}{(a_j + b_j)/2} \quad (5)$$

where a_j is the percentage of questions devoted to subject j in the examination (based on consideration of nonexcluded questions only) and b_j is the percentage of the course time devoted to subject j. This in turn would have resulted in the partial correlation coefficient

$$r_{(Q_{\cdot j},\, P_{\cdot j} | E_j)} \quad (6)$$

where the notation $r_{(1,2|3)}$ is used as synonymous with $r_{12.3}$ and $Q_{\cdot j}$ is the student rating of instruction in subject area j, being substantially lower than the correlation coefficient

$$r_{(Q_{\cdot j},\, P_{\cdot j})} \quad (7)$$

and presumably of no statistical significance. Reference to my article (*2*, p. 568) shows this was not the case.

Rodin also claims (second paragraph of her comment) that because of the confounding of instructor and subject matter, the positive correlation I observed between student ratings of instruction and class performance could have been spuriously raised by class performance (and presumably student ratings) being higher in subject areas the students liked better. In this respect it is crucial to consider when such a preference for some subject areas would have arisen. Since the vast majority of the students had no formal contact with the discipline before the start of the course, it seems unlikely that they should have had preexisting preferences for some of the subject matter. Moreover, if students were predisposed to like some subject areas more because of some intrinsic quality of the subject area (for example, its being less mathematical), then one would expect this to be equally true of both my sample and the nationwide sample. The contention that the two samples were rather similar is supported by the high value of the correlation coefficient 2; accordingly, such a predisposition would not affect class performance as calculated in Eq. 1. If, on the other hand, as a result of the instruction received in a given subject area, students came to like it better, studied it more, and therefore scored higher on it in the examination, that would be an entirely appropriate pedagogic strategy synonymous with effective instruction.

Rodin misrepresents me when she says that I "rejected student performance on departmental exams as a measure of student learning." Nowhere did I make such a statement. Departmental examinations obviously do measure student learning. I did discuss the "problems inherent in using class performance on internal examinations as a measure of the teaching effectiveness of the faculty" (*2*, p. 568). Judging from various comments I have received and remarks in the third and eighth paragraphs of Rodin's comment, however, the pertinent statements in the concluding paragraph of my article could have been more explicit. Thus some individuals, noting the apparent contrast of there being a significant positive correlation between student ratings and class performance on national examinations but no correlation of student ratings with class performance on institutional examinations, have speculated that possibly "the national exam more closely represents the students' expectations of course content " (*3*). Such speculation is unwarranted: because no measure of question difficulty was available for the questions in the institutional examination, class performance on any given subject area in these exams was computed as the mean percentage of the class answering questions in that subject area correctly, or specifically as

$$y_{\cdot j} = \frac{\sum_{i=1}^{n} y_{ij}}{n} \quad (8)$$

That the correlation coefficient between student ratings and this measure of class performance, that is

$$r_{(Q_{\cdot j},\, y_{\cdot j})} \quad (9)$$

proved not significant should not, therefore, be taken as a reflection on institutional examinations, but rather as an indication of the importance of allowing for variations in intrinsic question difficulty. In my article, I refer to the national examination as a normative one precisely because information regarding the performance of the nationwide sample made it possible to allow for question difficulty. An institutional examination could also be considered normative if the performances of several classes on it were known.

In summary, I find no merit in the specific criticisms Rodin makes of my study. In any comparison of these studies it should be remembered that my study dealt with faculty as primary teachers. By contrast, the Rodins dealt with teaching assistants in an auxiliary role, and accordingly their results might reflect, or be confounded by, an interaction between the primary teacher and the auxiliary instructors. In any event, their results should not be extrapolated to situations involving faculty as primary teachers.

PETER K. GESSNER
Department of Pharmacology and Therapeutics, Schools of Medicine and Dentistry, State University of New York, Buffalo 14214

References and Notes

1. M. Rodin and B. Rodin, *Science* **177**, 1164 (1972).
2. P. K. Gessner, *ibid.* **180**, 566 (1973).
3. W. A. Brown, personal communication.
4. I am grateful to W. J. Popiel, N. Solkoff, and T. Gessner for helpful comments and to W. E. Schotz, S. Addelman, and N. C. Severo for statistical advice.

23 September 1974