TO: President O'Leary  
Vice Presidents  
Deans  
Department Chairs  
Members, Council on Educational Policy

FROM: Patrick T. Terenzini

DATE: March 10, 1983

RE: Summary Report of the Pilot Test of the  
Student Instructional Rating Form (SIRF)

In March of 1980, the Council on Educational Policy (EPC) passed a resolution calling for a pilot test of the "Student Instructional Rating Form" (SIRF) which had been recommended by the Committee for the Improvement of Undergraduate Teaching (the Tompkins Committee) and subsequently modified by EPC and its Committee on Evaluation Policy. SIRF was designed at the request of EPC for potential use in personnel decision-making and is not to be confused with the Survey of General Education Courses used for the first time this past Fall and required by the University Senate bill establishing the General Education Program.

The pilot test was conducted in the Spring of 1981, and the attached summary describes the main findings and conclusions of that test. In brief, results indicate that the SIRF items comprise a highly reliable device for measuring students' opinions of the instruction they receive. So far as the validity of the measure is concerned, given the unavailability of the data needed to conduct conclusive studies, it can at least be said that the SIRF items perform in ways consistent with those of similar instruments in use on other campuses and/or commercially available.

A copy of the complete technical report (65 pp.) will be made available to each department upon request. A limited number of copies of the attached summary are available to interested faculty members, also on request. We will appreciate it if department heads will advise their faculty members of the availability of this summary.

We wish to thank publicly the faculty, students and staff of the following academic units for their patience and cooperation in the pilot test:

Anthropology  
Art  
Atmospheric Sciences  
Biological Sciences  
Classics  
Counseling Psychology and Student Development  
Economics  
Educational Admin. & Policy Studies  
English  
Geological Sciences  
Philosophy  
Physics  
Program Dev. & Evaluation  
Public Administration  
School of Business  
School of Criminal Justice  
School of Social Welfare  
Sociology

cc: H. Desfosses  
F. Lees  
P. Krosby  
L. Hartmark  
M. Rozett  
F. Pogue  
F. Volkwein
SUMMARY

REPORT OF THE PILOT TEST OF
THE STUDENT INSTRUCTIONAL RATING FORM (SIRF)

December 1982

Persons wishing more detailed information on the methods and analytical procedures
used in this study are invited to contact Patrick T. Terenzini, Director of
Institutional Research, Administration 260. Telephone 457-4621.
SUMMARY OF THE REPORT OF THE PILOT TEST OF 
THE STUDENT INSTRUCTIONAL RATING FORM (SIRF)

The pilot test of the Student Instructional Rating Form (SIRF) was conducted in the Spring semester of 1981 and involved eighteen departments, schools or colleges. Only sections with enrollments of ten or more students, and taught primarily as lectures or seminars were included. Analyses were based on data received from 367 sections meeting the inclusion criteria and constituting 77 percent of the sections eligible in the 18 departments and schools and nearly 30 percent of such sections taught in all departments campus-wide. While the pilot test participants were not randomly drawn, nor can any claims be made for their representativeness, the test was based on a large sample of students (over 9,500 usable responses) and instructors (245 individual instructors participated, 197, or 80%, of them members of the regular teaching faculty) in the 367 sections. And while the distributions of students, instructors and sections may depart from their respective population distributions by non-chance amounts, examination of the pilot sample distributions indicate considerable variety and substantial numbers in all categories of relevant dimensions (e.g., students' class status, sex, reason for taking the course, instructor's rank and type of appointment, course level and size). Moreover, the response rates within sections were generally high, the mean being 75 percent; less than 8 percent of the sections had response rates below 50 percent.
The analyses performed suggest a number of conclusions about the SIRF items. First, the ratings students assigned tended to cluster at the upper (positive) end of the scale. This result is not unexpected, although one can only speculate on the reason(s) behind it. One possible explanation is that the quality of instruction at Albany is, overall, quite high. Other, more generalized surveys of Albany students' opinions of the quality of the instruction they receive support this proposition. Another possible explanation is that the skewness is due, in part, to the directions students received when completing the SIRF forms. The form states that ratings "may be used in making promotion, tenure or other merit decisions concerning this instructor." (This will not, in fact, be the case with the present data, which will be used for such decision-making only upon request of the instructor.) The research suggests that students tend to rate instructors higher when the students know the ratings may be used in personnel decision-making than when they are told the results are intended for instructional improvement and will be seen only by the instructor.

Second, a series of factor analyses consistently indicated that the nine SIRF items describe a single dimension of teaching, one that might reasonably be characterized as a "skill" factor. That factor accommodated 58.4 percent of the total variance among the items at the individual student level of analysis, and 70.4 percent of the total variation among the section means. Students' ratings of their "Instructor, overall" had the highest loading on
the factor, followed in order by the instructor's ability to stimulate student interest in the course material, overall rating of the course, and the instructor's ability to challenge students intellectually.

Third, whether at the individual student or section level of analysis, the nine SIRF items appear to be reasonably strongly correlated. At the individual level, the item correlations ranged from .38 to .75, with a mean of .525. At the section level, the correlations ranged from .39 to .91, with a mean of .658. These analyses, however, together with the number of students not responding to each item, strongly indicate that items 5 ("Was receptive to students' ideas and viewpoints") and 6 ("Was available outside class to discuss course matters") may be the weakest of the set. These two items consistently had the lowest correlations with other items, the lowest factor loadings and communalities, and, in the case of item no. 6, over a fifth (21%) of the students were unable to respond (i.e., chose the "Don't know, or doesn't apply" response option). The correlations of these two items with section composite ratings (the sum of the section mean item ratings) were also the lowest. Whether these items should be retained in SIRF, however, is a policy, not empirical, issue that deserves discussion.

Fourth, while the test-retest reliability (stability of the ratings over time) cannot be assessed because of the unavailability of appropriate data, the internal consistency reliability of the composite scale (the degree to which the items
comprising the scale appear to measure the same trait) is quite high, ranging from .91 (at the individual level of analysis with all 9 items included) to .96 (at the section level of analysis with items 5 and 6 excluded). Calculations to determine the amount of error that might be present in any given instructor's rating in a single course indicate that an individual's "true" rating might be expected in 95 out of 100 testings to fall within an interval defined by .21 plus and minus the composite rating actually received. Moreover, analyses indicate that the upper and lower fifteen percent of all sections constitute reliably different groups. Put another way, if an individual's composite score lies at or below the fifteenth percentile, it can be said with 95 percent confidence that that instructor's "true" rating is not, in reality, in the 85th percentile or higher.

Students in the SIRF pilot sections can be considered random samples of all students who might have enrolled in those sections. Assessments were made of the dependability of these ratings for generalizing to potential future sections taught by the same instructor, and results indicate that dependability is quite high (with items fixed, only one coefficient below .90) across colleges and schools. Moreover, the reliability of scores across samples of students vary only slightly with section size—ratings are not substantially less reliable in smaller sections (12-15 students) than in larger ones.

Fifth, to the extent that the Tompkins Committee's review of available instruments and related materials can be said to be
adequate, and to the extent that the faculty members who sat on
the Tompkins Committee, EPC, UAC or the Committee on Evaluation
Policy and either developed or reviewed the nine SIRF items can be
considered experts in teaching, to that extent one might conclude
that the SIRF items have been content validated for their intended
purpose on the Albany campus. (Content validity is concerned with
the degree to which an instrument contains a representative
sampling from the behavioral domain to be measured.)

Sixth, the unavailability of the necessary data precludes
an assessment of the criterion-related validity of the SIRF form,
that is, the measure's correlation with some independent measure
of what the form is intended to assess. This type of validation
typically involves multi-section courses with a common final
examination and the correlation of mean instructor ratings with
the average grades earned by the instructor's students. Albany
offers no such courses. Such a validational assessment is,
clearly, possible, but it will require additional research and
extensive cooperation from departments and individual faculty
members.

Seventh, beyond the results of the factor analyses
described earlier, the construct validity of the SIRF form (the
extent to which it can be said to measure a theoretical construct
or trait) can be measured only indirectly, through assessment of
what is called its "convergent and discriminant validity." These
terms refer to the degree to which the SIRF items correlate with
things with which they should correlate (convergent validity) and
do not correlate with things with which they should not correlate (discriminant validity). Thus, given the research of others and intuition, one might expect the SIRF composite score to be related positively to students' class year (e.g., freshman, sophomore), reason for taking the course (e.g., required or elective), and whether the course is in the student's actual/planned major field. The results indicate that SIRF scores are, indeed, positively related to these student traits.

Many people believe, although the belief is generally unsupported in the research literature, that the grades students receive (or expect to receive) are strongly and positively related to the ratings students give an instructor, that an unscrupulous instructor can insure favorable ratings by giving high grades. In the SIRF pilot test, instructors' composite ratings correlated .04 with students' cumulative grade-point averages and .09 with the grades students expected to receive in the courses being rated. Thus, for this set of data based on the SIRF items, it appears that the ratings students give their instructors are virtually unrelated to the grades they expect or their cumulative grade-point averages.

While one might predict these relations to exist (or not exist), one might also expect the relation to be small. A multiple regression with all student traits above used as the predictor variables suggests that the strength of the influence of student traits on the ratings they assign is, indeed, slight. Student characteristics account for only 2.5 percent of the total
variation in the ratings assigned. Thus, it would seem reasonably safe to conclude that student traits (including expected grade) have so little influence on SIRF ratings that they may be educationally and administratively inconsequential.

The results of the analyses of the influence of instructor characteristics (e.g., rank, type of appointment) are similar to those for the student traits in that reliable and predictable relations were identified (e.g., senior faculty members and instructors with regular teaching appointments received higher ratings), but the magnitudes of the associations were negligible. When composite score was regressed on instructors' academic rank, type of appointment and sex, these variables—as a set—were found to explain only 4 percent of the total variation in composite ratings. As other research suggests and as one might hope, the strength of those relations is slight.

Analyses based on course or section characteristics, however, presented a somewhat different picture. Prior studies lead one to expect a negative relation between composite ratings and section size, and a positive correlation between ratings and section level of instruction. Both expectations are fulfilled in the SIRF data. Of the two, size was clearly the more important, explaining some 7 percent of the total variation in composite score ($r = -0.268$), indicating that as section size increased, ratings tended to decline. Similarly, a general linear trend was identified in the relation between section level and ratings, but the positive relation was slight, the two variables having less
than 2 percent of their variance in common.

Earlier studies have suggested that ratings may vary according to the discipline of the course being offered. This condition was found in the SIRF data, one college having significantly lower ratings than two other colleges. The absolute magnitudes of the differences were slight, however, and these were the only reliable differences among the set of six colleges.

When the three major course characteristics were regressed as a set on composite ratings, the set explained 16.5 percent of the total variation in the ratings (R = .381). These results clearly indicate the need to control the influence of course traits on ratings when the ratings of different instructors (or even ratings of the same instructor in different courses) are compared. Establishing SIRF form norms (comparison standards) based on course size, and probably course level and discipline, would seem highly desirable to guard against the misinterpretation of SIRF results.

Finally, all student, instructor and course traits, taken together, explained 22 percent of the variation in composite ratings (R = .469). Among the sets, however, course characteristics may account for as much as two-thirds of that 22 percent, again calling attention to itself as a set of variables that should be taken into account when SIRF results are interpreted.

By comparison with other such instruments (including commercially available ones), the nine SIRF items appear to
constitute a highly reliable device for measuring student opinions of their instruction. So far as the validity of the measure is concerned, and given the unavailability of the data needed to conduct conclusive studies, it can at least be said that the instrument performs in predictable ways—overall ratings correlated with things with which they should correlate, and did not correlate with things with which they should not. Moreover, even where statistically reliable relations between ratings and student, instructor or course traits were identified, those relations were negligible to modest in strength. When taken all together, these traits leave more than three-fourths of the total variance in ratings unexplained and potentially related to instructional skill.

One should not, of course, expect such characteristics to explain very large proportions of the variance in ratings: these traits are probably not the major determinants of teaching effectiveness. Similarly, one should not expect how much students learn (at least as measured by course grades) to be determined solely by the teaching competence of the instructor. Such an expectation would be a heavy burden for an instructor to bear, implying as it does that how much a student learns is independent of his/her level of motivation to learn, study habits, prior preparation in the subject field, academic aptitude and so on. The learning reflected in grades proceeds from many sources, some of which lie outside the classroom and an instructor's ability to influence. For the moment, however, the relation between student
opinions of their instruction (as measured by the SIRF items) and how much they learn is a matter for further research.

REFERENCES


Persons wishing more detailed information on the methods and analytical procedures used in this study are invited to contact Patrick T. Terenzini, Director of Institutional Research, Administration 260. Telephone 457-4621.
STUDENT INSTRUCTIONAL RATING FORM (SIRF)

Blacken the appropriate circle heavily and completely with a No. 2 pencil only. If you wish to change your response, erase the mark completely. Please make no marks in grey-shaded areas.

NOTE: Your responses to the first 9 items may be used in making promotion, tenure or other merit decisions concerning this instructor.

INSTRUCTIONS: Please code the 4-digit CALL NUMBER of this course in the unshaded area in the upper left corner of this sheet. Then, using the scale below, indicate how frequently each statement was characteristic of this instructor.


Blank - Don't Know/ Doesn't Apply

1. Was well prepared for class

2. Communicated course content in ways you understood

3. Stimulated your interest in the course material

4. Challenged you intellectually

5. Was receptive to students' ideas and viewpoints

6. Was available outside class to discuss course matters

7. Held you to high standards of performance

8. INSTRUCTOR overall: 5 = Excellent 4 = Good 3 = Average 2 = Fair 1 = Poor

9. COURSE overall: 5 = Excellent 4 = Good 3 = Average 2 = Fair 1 = Poor

10. Are you: 1 = Fr. 2 = Soph. 3 = Jr. 4 = Sr. 5 = Grad. 6 = Other

11. Is this course: 1 = Specifically required, 2 = Required but a choice among several; 3 = An elective

12. Is this course in your actual/planned Major = 1  Minor = 2  Other = 3

13. Is your current GPA closest to: 4.0 = 4  3.0 = 3  2.0 = 2  1.0 = 1

14. Expected course grade: A = 5  B = 4  C = 3  D = 2  E/U = 1  S = 6

15. Are you: 1 = Male 2 = Female

Thank you and please return on No. 2 pencils. (Someone other than the instructor should collect and return these forms)
GENERAL PURPOSE QUESTIONNAIRE

INSTRUCTIONS

Use a number 2 or softer lead pencil. Make all marks in the response circles. They should be dark and glossy, as shown below.

Do not make any stray marks. Erase completely if you change your mind. Mark any requested information on the reverse side.

INCORRECT MARKS

CORRECT MARKS
Early in Spring 1981, the following eighteen (18) colleges, schools and/or departments participated in the SIRP pilot test:

**College of Humanities and Fine Arts**
- Art
- Classics
- English
- Philosophy

**College of Science and Mathematics**
- Atmospheric Sciences
- Biological Sciences
- Geological Sciences
- Physics(a)

**College of Social and Behavioral Sciences**
- Anthropology
- Economics(a)
- Sociology

**School of Education**
- Educational Administration and Policy Studies(a)
- Counseling Psychology and Student Development
- Program Development and Evaluation

**Graduate School of Public Affairs**
- Public Administration

**School of Business(a)**

**School of Criminal Justice(a)**

**School of Social Welfare**

(a) Participation restricted to selected sections.
In late March, 1981, the heads and faculty of participating departments received a memorandum describing how the pilot test would be conducted, as well as the directions for administering the forms and for returning complete forms (see Appendix 1). Department heads also received a listing of the sections thought to be participating in the pilot and were asked to notify the Office of Institutional Research in the event that any section has been overlooked or erroneously included.

Unless advised otherwise by deans or department heads in the participating schools and departments, the pilot test population included all sections meeting both of the following criteria: a) had enrollments of 10 or more students (according to the Registrar's file as of April 13, 1981) and b) were taught primarily as either lecture or seminar sections according to the coding in the University's Course and Section Analysis (CASA) system. Independent study, laboratory, and discussion sections were excluded.

For each of these sections, a mailing label was produced containing the instructor's name, the course number, section number, the number of students enrolled, and the office address of the department offering the course. This mailing label was placed on a manilla envelope containing sufficient forms for the section's enrollment, as well as instructions for administering and returning the forms. A second label was affixed to each envelope and asked the instructor to verify the course information contained on the mailing label and to indicate (by signature) that
the course being rated was taught by the person named on the mailing label. One packet was prepared for each section with a discrete call number, and packets were delivered to participating departments approximately three weeks before the end of the semester.

Faculty members were advised that the forms could be completed during a class meeting of their choice, but they were discouraged from administering the forms either immediately before or after giving an examination. Instructors were asked to leave the room while the forms were completed and to designate a student to return the package of completed forms to any of a number of drop points around the academic podium (a map indicating drop points was included in each packet). Many departments assigned administration of the forms to graduate assistants. Returned forms were checked by Office of Institutional Research staff for proper coding of call number, and uncompleted forms were removed from the packets. Packets were then scanned by the optical reading equipment in the Computing Center, and the data were prepared for analysis and reporting.