A LIVELY ELECTRONIC COMPENDIUM OF RESEARCH, NEWS, RESOURCES, AND OPINION

Astronomy Education Review

Volume 4, Oct 2005 - Jul 2006
Issue 2

Assessment of Large General Education Astronomy Classes

by **Thomas H. Robertson**
Ball State University Department of Physics and Astronomy
W. Holmes Finch
Ball State University Department of Educational Psychology
Received: 05/26/05, Revised: 12/22/05, Posted: 01/31/06

The Astronomy Education Review, Issue 2, Volume 4:28-53, 2006

Abstract

We present results for a decade-long assessment program for an introductory course for non-science majors. This study focuses on student evaluation data and student-supplied information about sex, race, age, academic background, student study time, study habits, and course grade. The results reveal a statistically significant positive relationship between the student evaluation means and grade inflation, and negative relationships between student evaluation means and student study time and between student evaluation means and previous course performance in chemistry and physics. The positive relationship between student evaluations and grade inflation was not surprising. However, the negative relationship between student evaluation means and previous course performance in chemistry and physics was unexpected. Course grades are significantly positively related with estimated grade point average (GPA) and previous course performance in chemistry and physics, and negatively correlated with student study time. Positive associations between course grades and GPA and previous course performance in chemistry and physics were expected. The negative relationship between course grades and study time was unexpected, but reports of similar results have been found in the literature. Course grades of male students were significantly higher than those of female students. We discuss the implications of these results for personnel and program evaluation.

1. INTRODUCTION AND BACKGROUND

The use of student ratings for personnel decisions has been debated for decades. Hake (2002) advocated the use of specifically developed examinations to measure learning and suggested that student ratings do not provide reliable evidence of learning. Cashin (1988, 1995) argued that a variety of studies demonstrate that student ratings are valid measures of student learning. Most of Cashin's validity arguments relate to

correlations of student ratings determined by evaluation techniques that may be equally inadequate to measure student learning. Although most faculty members are skeptical about the validity of student ratings as measures of teaching effectiveness, such ratings are frequently required of persons applying for promotion, tenure, or merit pay.

Of special interest in Cashin's conclusions are the biases in student ratings, which he did acknowledge. Student ratings show bias in (1) student motivation, (2) subject area, (3) course level, and (4) course grades. Students enrolled in courses voluntarily rated them higher than students enrolled in courses required for programs such as general education. Students enrolled in humanities and arts courses rated them higher than students enrolled in social science courses, and students enrolled in social science courses rated them higher than students enrolled in math and science courses. Students enrolled in higher-level courses produced ratings higher than students enrolled in lower-level courses. Instructors in the Astro 100 course, the subject of this study, were negatively impacted by all three of these biases. Finally, students expecting to receive higher grades produced ratings higher than students expecting to receive lower grades.

Although Cashin recommended controlling for these biases, no mechanism is in place within our university to do so. Student ratings are required of all faculty members being considered for promotion and tenure, but information about these biases is not provided to committee members evaluating the credentials of faculty members. And although our department requires all faculty members to provide grade distribution data with each student rating report, such information is not required at the college level, where final promotion decisions are made. The average data provided by our internal assessment program do permit candidates to provide mean data for the multiple sections for this course in order to permit a more informed evaluation of their ratings by others.

Student evaluations of instructors and courses are required in many colleges and universities and are often considered in making personnel decisions such as tenure and promotion. They can also be used as evidence of curricular effectiveness in program evaluation. Unfortunately, these applications often do not distinguish between small and large class size, freshman and upper-division courses, and courses taken by majors and those completed as part of required curricula. This project was undertaken initially to provide a context for the interpretation of student evaluation data within our physics and astronomy department. Courses within the department are divided into five different major groups, and student evaluations for individual faculty members are compared with those of other faculty members teaching similar courses. The five groups are (1) general education, (2) teacher education, (3) algebra-based physics, (4) calculus-based physics, and (5) all advanced physics courses. Information provided within each of these groups is much more useful to faculty members and administrators than simple departmental averages.

This article presents the information collected on these evaluation forms for our Astro 100 classes over a 10-year period, from 1993 to 2002. Astro 100 is a traditional astronomy survey course for non-science majors. It is one of several physical science courses that students can take to fulfill distribution requirements in the Ball State University general education program, the University Core Curriculum. Although some of these courses fulfill science requirements for physics majors and minors and other science curricula, Astro 100 was designed specifically for this program. The course was taught in sections of about 100 students each during the academic year, with smaller sections offered in the summer and off-campus. This course was taught in the lecture-discussion mode during this period, with the amount of discussion varying with the instructor assigned to each section. Each instructor had flexibility in selecting the specific topics to be covered in his or her section. Few instructors covered the entire book, and

considerable differences in emphasis existed from instructor to instructor. Three observational astronomers and several physicists taught sections of this course during this period. A hard-wired student response system was installed in the classroom used for this course in 1997. Although most of the Astro 100 instructors learned to use this system, it was employed primarily for taking attendance and administering quizzes and exams. Few instructors used the system to modify their courses to include surveys of students and discussion of concepts involving peer instruction. After the first year of use, the keypads started to become unreliable, and use of the system declined in subsequent years. Thus, the system would have had little impact on the results of this project, which sampled students in 1996 and 1999.

The data for Astro 100 were analyzed to identify correlations of mean student evaluation scores for the instructor, course, and course goals with sex, race, age, academic background, student study time, study habits, and grade inflation. Student course grade data were analyzed to identify correlations with the same factors, with the exception that estimated GPA was used in place of grade inflation as an independent variable.

2. STUDENT SURVEY FORMS

Student survey data were collected on short and long forms. The short forms consisted of 11 specific instructor-related items, 4 specific items related to the course in general, 1 overall instructor item, and 1 overall course item. The evaluation items were constructed as positive statements, with student responses ranging from 1 (*strongly agree*) to 5 (*strongly disagree*). Thus, higher values reflect poorer ratings. It might be helpful to think of the system as a ranking system rather than a rating system. Having dealt with magnitude systems for 2,000 years, astronomers should be familiar with this concept. Short-form evaluations were completed in nearly all courses at the end of the fall semester and as desired by the instructor in other terms. The long forms used to collect student data contain these same items and academic background data, demographic data, information about student study time, self-evaluation of study habits, estimated GPA, and expected course grade. Such information was requested (1) to determine whether correlations existed between these different variables and student evaluations of the instructor and course and (2) to attempt to identify correlations with course performance as reflected by the expected grade. It is important to understand that all information was provided by the student and was limited by student honesty and memory. The evaluation forms were administered by a department faculty member other than the course instructor and were completed in class. It typically took 10 minutes to complete the short-form evaluation and 20–30 minutes to complete the long-form evaluation. Students who did not attend class regularly during the last few weeks of the term are not represented.

3. DATA COLLECTION AND USAGE

Ball State University has offered an introductory astronomy course as part of a general education program for almost 20 years. Annual enrollments in this course have ranged from 1,600 to 2,000 students per year. Most students are enrolled in sections with a limit of about 100 students. Since 1993, the long evaluation form described above has been administered during the last few weeks of all fall semester classes once every three years. Data have been collected in 1993, 1996, 1999, and 2002. These data form the basis for this analysis. About seven faculty members teach eight to ten sections of this course each fall. Nine different faculty members have taught sections of this course during these assessment years. The number of student evaluations each year ranges from 466 to 618. Typical attendance during the final two weeks of the semester has resulted in completion rates of 60%–70% on evaluations. In terms in which long-form

data are collected, evaluation forms for each section of Astro 100 were processed separately. A section mean value was computed for each item, and a mean was computed across all sections to determine a course mean. In addition, all evaluation forms were processed in a single batch to compute mean values as well. Each instructor was provided with his or her personal section mean results and the means for all sections and forms. These data permitted instructors to identify personal strengths and weaknesses. Similar data were used by the department chairperson to complete teaching effectiveness reviews for all faculty. These reviews were used to make recommendations pertaining to merit pay, tenure, and promotion.

We expect that reliance on student-supplied data will limit the reliability of information, but self-reported data are better than no data at all. Cassady (2001) found that self-reported GPAs can be very reliable in assessment studies. The responses to questions regarding previous success in chemistry and physics courses are probably not as reliable because they are more complex and rely upon students recalling courses that they may have taken several years before the survey. The student study time data are probably the most subjective.

4. METHODS

Traditional student evaluations provide information that can be used to compute the average responses of students in a given section or course. However, they do not permit analyses of the characteristics of the students submitting the evaluations. What is the academic background of a student? What is the general level of academic ability of a student? How is the student performing in the course? Is the expected grade in the course for this student higher or lower than his or her overall GPA? Are there significant correlations between these variables and the evaluations of the students? Are there significant correlations between these variables and the expected grades of the students? The assessment program developed in the Department of Physics and Astronomy was designed to collect data that could be used to answer such questions. The program was established initially to avoid any connection between the student evaluation form and the official record of the student. Although this assessment program may have decreased reliability of the data somewhat, it significantly simplified the collection of data and removed any concern on the part of the student that his or her evaluation may be traced to him or her at some future time, and may therefore have encouraged greater honesty in rating various aspects of the course.

Initial analyses of the data included use of multiple-linear regression models for the variables, which can be represented by numbers on a continuous scale (levels of agreement, course grade, GPA, amount of study time) followed by student's *t*-tests for categorical variables such as sex and race. This model would produce an equation of the form:

Overall Instructor Rating =

C_1 * Grade Inflation + C_2 * Chem/Phys Background + C_3 * Study Time This Course

The *overall instructor rating* is the dependent variable in this model. A variety of independent variables were included in this analysis, but only those that are statistically significant would be included in the final model shown above. The coefficients C_1, C_2, and C_3 are the slopes of lines showing the linear relationship between the respective independent variable and the dependent variable. Following the construction of such models, it is possible to identify reasons for differences between means in the dependent variable for different categories. For example, a simple student's *t*-test may reveal that male students evaluate an instructor more favorably than do female students. If female students have

significantly different chemistry/physics backgrounds, this would explain some of the difference. If male students receive systematically higher grades, this would also be revealed. The coefficients permit the calculation of the change in the dependent variable, which can be attributed to a given difference in an independent variable.

The survey items and the definition and construction of model variables are described in the remainder of this section. Section 5, Results, contains tables and figures that show the dependent and independent variables and coefficients for the significant independent variables.

The term *grade inflation* in this article has a very specific definition: it is the difference between the grade expected in this course (as reported by the student) and the estimated GPA (as reported by the student). Thus, though the term used is *inflation*, it is possible for students to report grade *deflation* as well; that is, the expected grade in the course could be lower than the estimated GPA. Cashin's discussion of biases in student ratings clearly delineates the different interpretations of this variable. Some suggest that it reflects a bias in ratings whereby unwarranted assignment of higher grades is correlated with more positive student ratings. Others argue that better teachers produce better students and that the higher grades reflect real differences in instructor effectiveness. Because student learning is not measured by some objective means and as long as grades are assigned by subjective criteria that vary from instructor to instructor, it is impossible to resolve the relative merits of these two arguments. The choice of this variable name is not meant to imply that arbitrary assignment of higher grades is the primary factor or that increased student learning does not play a role. We do believe that this variable is more meaningful than the expected course grade.

The hierarchical linear models (HLM) method of analysis allows for the modeling of a dependent variable as a function of a set of independent variables of different types. A thorough discussion of this methodology can be found in Bryk and Raudenbush (1992), while examples of its application to educational data appear in Gustafson and Hendel (2004) and Grodsky and Gamoran (2003), for example. The results of this analysis include statistical tests for the variables of primary interest and measures of the impact of other effects that are anticipated to influence the dependent variables. To ascertain the nature of the relationship between the dependent variable and the continuous (as opposed to categorical) independent variables, slopes are estimated for each covariate that is found to be statistically significant. Statistically significant results for categorical effects are followed up by the Tukey-Kramer post hoc test to determine which categories have significantly different means. For example, the test allows for the comparison of the mean scores of freshmen with sophomores, freshmen with juniors, and so on.

To address the research questions of interest, a set of HLM was used, treating as dependent variables the mean of several instructor ratings items, the overall instructor rating, the overall ratings of course content, the overall ratings of course design, and expected grade. Composite variables were obtained by averaging responses to specific groups of items on the course evaluation, which appears in the appendix. Specifically, for an individual student, the instructor mean is composed of the mean of items 1–11, the course design rating is calculated from the mean of items 12–15, and the course content is calculated from the mean of items 16–20. These are summarized in Table 1. The overall instructor rating is given by the single item 21 (not in the table): "This instructor was one of the better University Core Curriculum instructors I have had at Ball State University."

Table 1. Construction of Composite-Dependent Evaluation Variables from Survey Items

Instructor Mean	1. My instructor displays a clear understanding of course topics.
	2. My instructor seems well-prepared for class.
	3. My instructor emphasizes conceptual understanding of the material.
	4. My instructor is careful and precise when answering questions.
	5. Exams cover a reasonable amount of material.
	6. Exams are fair.
	7. My course grade accurately reflects my knowledge of the material.
	8. The grading system was clearly explained.
	9. The grading system in this course is appropriate.
	10. My instructor returned graded materials in a timely manner.
	11. The teaching strategy used in this course was appropriate.
Course Design	12. I am generally pleased with the textbook in this course.
	13. The assigned reading is well-integrated into the course.
	14. The amount of material covered in the course is reasonable.
	15. The workload in this course is appropriate.

Course Content	16. This course was effective in improving my understanding of the universe.
	17. This course was effective in improving my understanding of natural laws.
	18. This course was effective in improving my understanding of the history of science and its relationship to human civilization.
	19. This course was effective in improving my awareness of the space program and its impact on mankind.
	20. This course was effective in improving my understanding of the nature of science in general.

Several independent variables are also composites of items taken from the course evaluation questionnaire as noted in Table 2.

Table 2. Construction of Composite Independent Variables from Survey Items

Study Time This Course	43. About how much total time did you normally spend preparing for this course from Monday through Friday each week?
	44. About how much total time did you normally spend preparing for this course on Saturday and Sunday each week?
	45. About how much total time did you normally spend specifically preparing for each hour exam in this course, in addition to normal preparation?
Study Time All Courses	46. About how much total time did you normally spend preparing for all courses from Monday through Friday each week?
	47. About how much total time did you normally spend preparing for all courses on Saturday and Sunday each week?
	48. About how much total time did you normally spend specifically preparing for each hour exam in any course in addition to normal preparation?

Previous Chemistry/Physics	32. How many semesters of high school chemistry have you completed with a grade of C or above?
	33. How many semesters of high school physics have you completed with a grade of C or above?
	37. How many semesters of college chemistry have you completed with a grade of C or above?
	38. How many semesters of college physics have you completed with a grade of C or above?
Overall Study Habits	49. How often during the course did you read the assignments before attending class?
	50. How frequently during the course did you take notes in class?
	51. How frequently during the course did you review your notes before the next class?
	52. How often during the course did you study with someone else in the class?
	53. If study guides or learning objectives were provided by the instructor, how frequently did you consult them?

For each of the independent variables in Table 2, the final value is calculated as the mean of the ratings provided by the constituent items. For example, the value for *overall study habits* is obtained for an individual student by averaging their responses to items 49, 50, 51, 52, and 53. Finally, the difference between students' performance in the astronomy course and their performance in college overall is represented as their expected grade in the course and their overall university GPA, where an A = 4, B = 3, C = 2, D = 1, and F = 0 for both variables. Positive values of this grade inflation indicate that students expected to do better in the class than in college as a whole, while a negative sign suggests the opposite.

The categorical demographic variables included in this analysis are student sex, class in school (freshman, sophomore, and so on), student race (white/Asian, all others), and student age category (younger than 23, 23–30, older than 30). Finally, the categorical variables that are treated as random effects in the HLM analysis include year (1993, 1996, 1999, and 2002), course section nested in year and teacher, and teacher.

5. RESULTS

Results for the course and instructor satisfaction ratings given by students will be addressed first, followed by the self-reported course performance data.

5.1 Course/Instructor Satisfaction: Student Evaluations

The distributions for the demographic variables used in this study, compiled across the four academic terms, are presented in Table 3.

Table 3. Demographics of Sample: Percent, Mean (SE)

	Percent: all years	Past success chem/phys*	Study time this course*	Study time all courses*	GPA*
Race					
White/Asian	92.32%	4.23 (0.05)	2.09 (0.02)	3.15 (0.02)	2.78 (0.02)
Others	7.68%	4.09 (0.16)	2.30 (0.07)	2.98 (0.09)	2.49 (0.06)
Class					
Freshman	47.15%	4.34 (0.06)	2.06 (0.02)	2.99 (0.03)	2.50 (0.04)
Sophomore	34.17%	3.99 (0.07)	2.14 (0.03)	3.22 (0.04)	2.63 (0.03)
Junior	11.25%	4.15 (0.12)	2.18 (0.05)	3.30 (0.07)	2.90 (0.04)
Senior	6.71%	4.75 (0.20)	2.03 (0.06)	3.41 (0.09)	3.00 (0.05)
Other	0.72%	4.40 (0.34)	2.00 (0.16)	3.56 (0.27)	2.97 (0.25)
Sex					
Female	45.52%	4.15 (0.06)	2.22 (0.03)	3.37 (0.03)	2.70 (0.03)
Male	54.48%	4.27 (0.06)	2.00 (0.02)	2.95 (0.03)	2.65 (0.03)
Age					
< 23	87.72%	4.29 (0.04)	2.06 (0.02)	3.12 (0.02)	2.69 (0.02)
23–30	8.21%	3.86 (0.15)	2.18 (0.06)	3.20 (0.08)	2.32 (0.08)
30+	4.08%	3.65 (0.21)	2.78 (0.11)	3.44 (0.12)	2.93 (0.07)

*Calculated for individuals as the mean of selected items in Table 2. Higher values indicate greater past success in chemistry/physics, more time spent studying for the course, more time spent studying for all courses, and higher self-reported GPA.

Means and standard errors for the dependent variables by the independent variables appear in Table 4. The mean values and error bars for two standard errors are shown in Figure 1 for the *instructor mean* and Figure 2 for the *overall instructor* variables, respectively. It should be noted that the distribution of these variables has been fairly consistent over time, thus allowing for their presentation in this unified format.

Table 4. Means (Standard Errors) of Dependent Variables by Levels of Independent Variables

	Instructor mean*	Overall instructor	Course design*	Course content*
Race				
White/Asian	1.91 (0.02)	2.48 (0.03)	2.19 (0.02)	2.26 (0.02)
Others	2.01 (0.06)	2.43 (0.10)	2.18 (0.06)	2.28 (0.08)
Class				
Freshman	1.86 (0.02)	2.36 (0.04)	2.15 (0.02)	2.42 (0.03)
Sophomore	1.93 (0.02)	2.52 (0.05)	2.21 (0.03)	2.25 (0.03)
Junior	2.01 (0.04)	2.63 (0.09)	2.27 (0.05)	2.27 (0.05)
Senior	2.06 (0.07)	2.74 (0.11)	2.28 (0.06)	2.41 (0.07)
Other	1.85 (0.23)	2.53 (0.38)	2.23 (0.29)	2.11 (0.30)
Sex				
Female	1.96 (0.02)	2.59 (0.04)	2.20 (0.02)	2.35 (0.03)
Male	1.87 (0.02)	2.37 (0.04)	2.18 (0.02)	2.19 (0.02)
Age				
< 23	1.93 (0.02)	2.50 (0.03)	2.21 (0.02)	2.83 (0.02)
23–30	1.83 (0.05)	2.33 (0.09)	2.07 (0.05)	2.16 (0.06)
30+	1.84 (0.09)	2.05 (0.13)	2.09 (0.09)	2.04 (0.10)

*Calculated for individuals as the mean of selected items appearing in Table 1. Smaller values indicate higher levels of satisfaction with that aspect of the course.

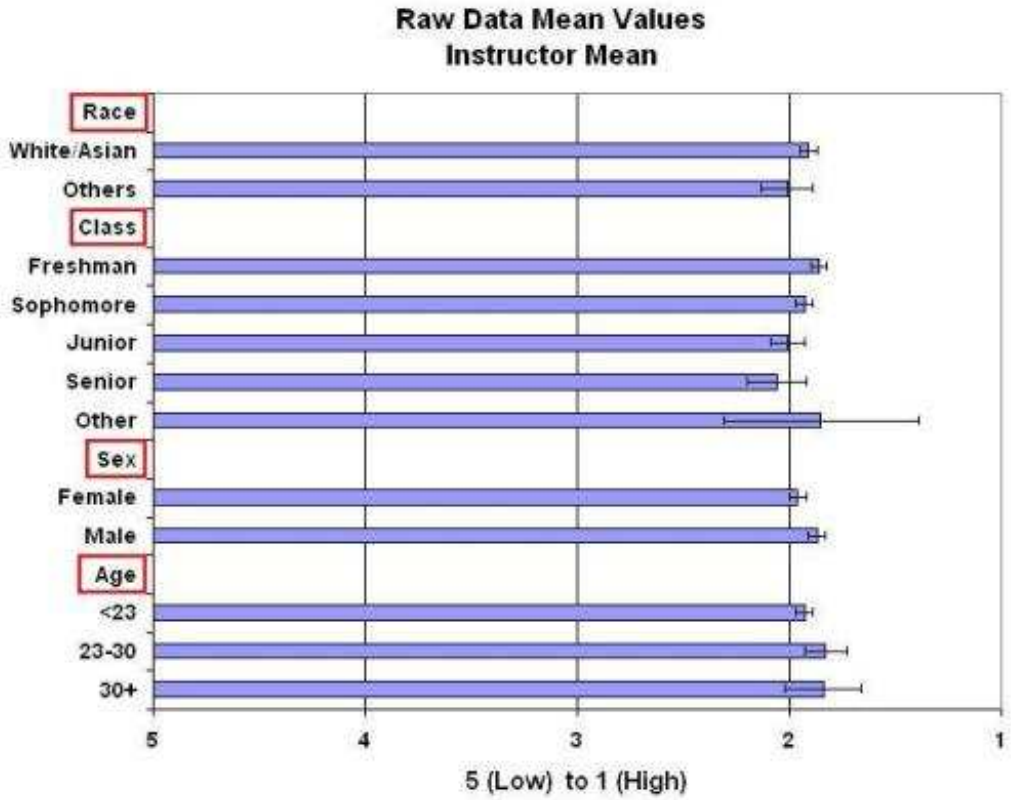

Figure 1. Raw Data Mean Values for Instructor Mean with Two-Sigma Error Bars

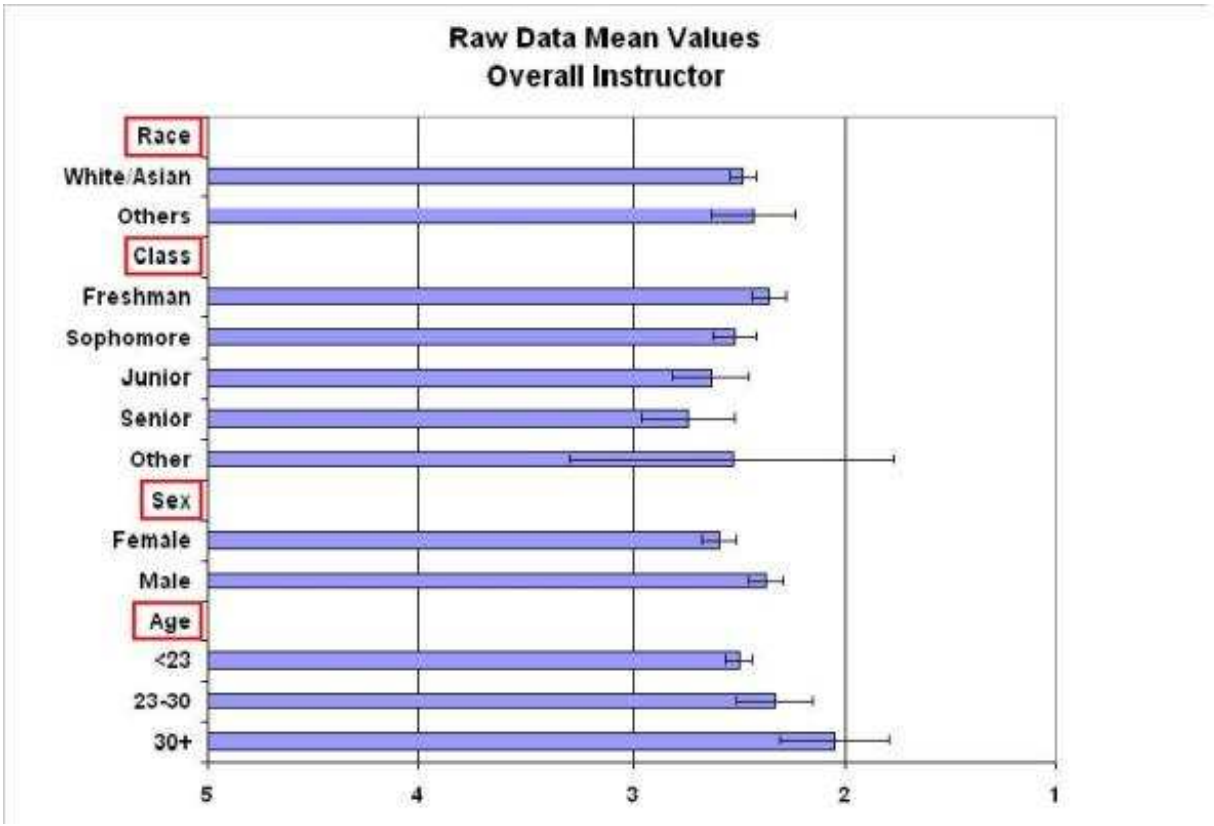

Figure 2. Raw Data Mean Values for Overall Instructor with Two-Sigma Error Bars

---

The contrast in content between the *instructor mean* and the *overall instructor* items is worthy of note. Most evaluation efforts report means or medians for individual specific items. Few compute mean values for these items. The data in Table 2, partially displayed in Figures 1 and 2, clearly show substantial, systematic differences between the mean of the 11 specific items (*instructor mean*) and the *overall instructor* item. The standard errors are also systematically smaller for the *instructor mean*, which is based on a set of specific items, selected by the faculty, that reflect faculty values for course instructors. The *overall instructor* item, on the other hand, solicits value judgments from students without reference to specific traits and may therefore be more subjective. In comparisons of student evaluation data for personnel decisions, it is expedient to view the *overall instructor* item as a summary of instructor effectiveness and to ignore the individual specific-item responses. These data suggest that the calculation of an *instructor mean* item can provide more objective data that are useful in evaluating teaching effectiveness.

The results of the significance testing for the HLM analyses for each of the rating variables appear in Table 5 (fixed effects) and Table 7 (random effects). The fixed effects are reported with the value of the *F* statistic and accompanying *p* value for each, while the random effects are reported both in terms of the variance component (the raw amount of variance accounted for by the independent variable) and the ratio of this component to the total variance in the dependent variable remaining after the fixed effects are

accounted for. The *F* statistic is the ratio of the amount of variation in the dependent variable due to the effect of the independent variable versus the amount of variation due to between-student differences. In general, larger values of *F* indicate that the independent variable has a greater impact on the value of the dependent variable than do simple between-student differences. Associated with each *F* is a probability, or *p* value, of obtaining this value of *F* if the independent variable has no impact on the dependent variable. Therefore, low *p* values indicate that it is unlikely that the independent variable has no effect on the dependent variable. Generally speaking, *p* values less than .05 are considered evidence of a significant effect. Another way to think about the *p* values is as the probability of concluding that an independent variable has a significant impact on a dependent variable based on the sample data, whereas in reality, this is not the case in the general population. Reaching such a conclusion is known as a Type I error.

In this study, there are five statistical models, one associated with each dependent variable (four course ratings and the expected grade). When more than one statistical test is conducted for a given independent variable using a single sample, the probability of making a Type I error at least once—concluding that one of the independent variables is significantly related to the dependent when in fact it is not—is greater than the standard of .05. A simple analogy can illustrate this point. If one plays a game involving flipping a fair coin in which victory is achieved every time heads appears, then the probability of winning in any one game is 0.5. However, the probability of winning at least once in 10 such games is much greater than 0.5. The same principles of probability hold in the case of making a Type I error. To control for inflation of the Type I error rate that is due to the use of five analyses (four course ratings and expected grade), a common correction to the *p* value, known as the Bonferroni correction, was employed in this study. This correction requires that the standard criteria for significance, .05, be divided by the number of analyses that are done—in this case, 5. The resulting value serves as the new Type I error rate for each of the individual analyses. Thus, in this case, the *p* value for each test is compared with 0.01 or 0.05/5.

Table 5. Results of Significance Tests for Fixed Effects

Variable	Instructor mean F(p)	Overall instructor F(p)	Course design F(p)	Course content F(p)
Student race	0.32 (0.57)	0.10 (0.75)	0.47 (0.49)	0.06 (0.81)
Class	1.87 (0.11)	2.19 (0.07)	2.53 (0.04)	1.90 (0.11)
Sex	0.00 (0.98)	0.38 (0.54)	1.73 (0.19)	3.19 (0.07)
Age	4.15 (0.02)	***6.28 (0.001)***	3.54 (0.03)	2.40 (0.10)
Study time this course	***7.36 (0.01)***	2.61 (0.11)	2.40 (0.12)	***6.45 (0.01)***
Study time all courses	0.70 (0.40)	0.01 (0.91)	0.00 (0.98)	0.24 (0.62)
Study habits	2.98 (0.08)	5.63 (0.02)	***9.83 (0.002)***	25.14 (< 0.001)
Chem/physics	***8.30 (0.004)***	***11.81 (< 0.001)***	4.08 (0.04)	***17.29 (< 0.001)***
Grade inflation	***59.95 (< 0.001)***	***58.35 (< 0.001)***	***43.29 (< 0.001)***	***40.10 (< 0.001)***

Based on these results, it appears that the amount of time spent studying for this course, the level of previous success in chemistry and physics, and the difference between their expected grade and their self-reported GPA (grade inflation) are all significantly related to the average of the instructor ratings. Similar results are evident for *overall instructor*, though the amount of time studying for this course is not significantly related to the item response, and there are differences due to age. With respect to the course design, overall study habits and grade inflation are statistically significant, while for course content, the amount of effort spent studying for this course, overall study habits, previous exposure to chemistry and physics, and grade inflation are significant.

To gain further understanding of the significant results, the slopes relating the covariates (continuous independent variables) to each of the dependent variables appear in Table 6.

Table 6. Slopes for Covariates by Dependent Variable

Variable	Instructor mean	Overall instructor	Course design	Course content
Study time this course	0.078	NS	NS	0.091
Study time all courses	NS	NS	NS	NS
Study habits	NS	NS	0.106	0.187
Chem/physics	0.032	0.071	NS	0.053
Grade inflation	–0.150	–0.278	–0.144	–0.152
Note: A discussion of the calculation of the variables appears in Tables 1 and 2. (NS = not significant)				

In this case, higher scores on each of the rating scales indicate a more negative impression of the teacher or course. The quantification of the evaluation items using positive statements with student responses ranging from 1 (*strongly agree*) to 5 (*strongly disagree*) results in negative slopes for positive correlations and positive slopes for negative correlations. Therefore, positive slopes for *study time this course* with instructor and course content ratings mean that the more time students spend studying for this course, the less they like it. Similarly, the better their study habits, the less students like the design of the course or the course content, and the more success students have had with chemistry and physics prior to taking this course, the less they like the instructor (using both measures) or the course content. On the other hand, the more positive the difference between their self-reported GPA and the grade they expect to receive, the more they like the instructor, course design, and course content. This suggests that when students expect to do better in the course than they have done in general, they rate the course more positively.

A follow-up analysis examining the differences in *overall instructor* ratings among the ages found that individuals under 23 years of age had significantly higher mean ratings than did the other two age groups. This result implies that younger students are generally less satisfied with the course than are their older counterparts.

Table 7. Results of Variance Components Analysis for Random Effects

Variable	Instructor mean component/ratio	Overall instructor component/ratio	Course design component/ratio	Course content component/ratio
Term	0 / 0	0 / 0	0 / 0	0 / 0
Teacher	0.096 / 0.219	0.305 / 0.202	0.005 / 0.013	0.093 / 0.153
Section	0.013 / 0.030	0.047 / 0.031	0.001 / 0.003	0.008 / 0.013
Residual	0.328 / 0.751	1.160 / 0.767	0.4281 / 0.984	0.508 / 0.833

In terms of the random effects, after the main effects are modeled, only the teacher accounts for more than 10% of the variation in *instructor mean*, *overall instructor*, and *course content*. In all cases, the residual accounts for the vast majority of variation in student ratings. This result suggests that once factors—such as the amount of time spent studying, prior success in chemistry and physics, and demographic information pertinent to students—are controlled, much of the score value is due to factors that are idiosyncratic to individual students rather than variables such as instructor, year in which the course was taken, or course section in a given semester.

The results of the analyses for the *overall instructor* rating item are similar to those obtained for the average of the 11 instructor specific questions (*instructor mean*). Specifically, in both cases, the age groups differ in the same fashion (i.e., younger students are less satisfied), and there is a statistically significant relationship between previous course work in chemistry/physics and overall satisfaction with the instructor. In addition, for both measures, the greater the grade inflation, the more satisfied students are with the instructor. The only difference in the models for the two types of instructor satisfaction is with the variable study for this course. It was found that this variable is related to the average of the instructor-related items but not to the overall rating of instructor.

Some insight into the relationship between responses to the overall instructor satisfaction item and attitudes toward various components of the course can be gained by using the correlation between the instructor, course design, and course content. The results appear in Table 8. The correlation coefficient, which appears in this table, reflects both the nature and degree of relationship between pairs of variables. It ranges in value between –1 and +1, with negative numbers indicating that larger values of one of the variables are associated with smaller values of the other. Conversely, positive correlation coefficients mean that larger values of one variable are associated with larger values of the other. It appears that *overall instructor satisfaction* is related to all three variables, though most with the *instructor mean*, as might be expected.

Table 8. Correlation between *Overall Instructor* and Average of Items for Satisfaction with Instructor, Course Content, and Course Design

Variable	Correlation
Instructor mean rating	0.70
Course content	0.63
Course design	0.51

5.2 Course Performance: Expected Course Grade

The means and standard errors of expected course grade for the various demographic groups included in this study appear in Table 9.

Table 9. Means of Expected Course Grade (SE) by Levels of the Independent Variables

Variable	Arithmetic Average
Race	
White/Asian	2.86 (0.02)
Others	2.62 (0.07)
Class	
Freshman	2.82 (0.03)
Sophomore	2.83 (0.03)
Junior	2.82 (0.06)
Senior	2.95 (0.08)
Other	3.33 (0.23)
Sex	
Female	2.74 (0.03)
Male	2.93 (0.02)
Age	
<23	2.83 (0.02)
23–30	2.83 (0.07)
30+	3.03 (0.10)

Mean values from Table 9 are displayed with two-sigma error bars in Figure 3.

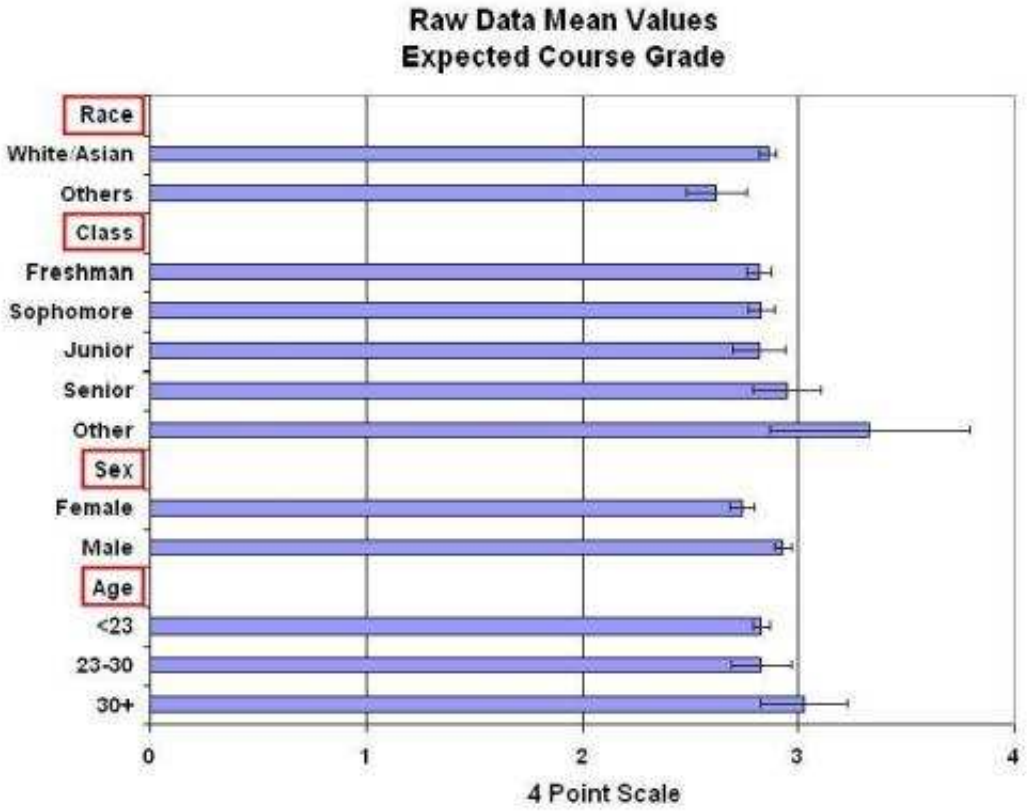

Figure 3. Raw Data Mean Values for Expected Course Grade with Two-Sigma Error Bars

---

In addition to examining the relationship of the demographic variables of interest here with course satisfaction, it is also of interest to ascertain how these factors are related to the grade that students expect to receive in the course. The reader should keep in mind that these are self-reports of the expected grade and not the actual grade that the students received in the course. Table 10 displays the results of the statistical tests for each of the factors included in this research, while Table 11 includes the results of the variance components analysis for this variable. One of the surprising results from the original model was a negative correlation between student study time and expected course grades. Because it seems unlikely that increasing study time will cause students to earn lower grades, the course performance model was repeated without *study time this course* as one of the independent variables. The results of this model are provided in Tables 10 and 11 for comparison with the original model. The most significant differences between these two models are in the *F* factors for *student race* and *study time all courses*.

Table 10. HLM Results for Expected Course Grade and Variance Components

Variable	Expected grade F(p) including *study time this course*	Expected grade F(p) without *study time this course*
Student race	2.19 (0.14)	2.87 (0.09)
Class	1.75 (0.14)	1.78 (0.13)
Sex	***15.62 (< 0.001)***	***16.09 (< 0.001)***
Age	6.27 (0.002)	4.87 (0.008)
Study time this course	***10.5 (0.001)***	(Omitted from model)
Study time all courses	3.27 (0.07)	.08 (0.77)
Study habits	0.45 (0.50)	.48 (0.49)
Chem/physics	***16.94 (< 0.001)***	***17.55 (< 0.001)***
Overall GPA	***110.64 (< 0.001)***	***111.05 (< 0.001)***

Table 11. Variance Components Analysis for Expected Course Grade

Variance components	Component/ratio including *study time this course*	Component/ratio without *study time this course*
Term	0.02 / 0.035	0.02 / 0.035
Teacher	0.001 / 0.002	0.001 / 0.002
Section	0.008 / 0.014	0.009 / 0.016
Residual	0.543 / 0.949	0.547 / 0.948

Given these results, it appears that the variables related to expected course grade include sex, age (partial), amount of time studying for all courses, previous exposure to chemistry and physics, and estimated GPA. The random effects in this analysis include the term in which the course was taken, the teacher, and course section. As with the results for course ratings described above, the residual effect can be thought of as

representing the impact of the students themselves on expected course grade. The second number in each cell of Table 11 (proportion of variation in dependent variable value due to the random effect) provides the most useful information about the relative impact of each variable. Based on these results, it is possible to conclude that none of the random variables accounted for more than 10% of the variation in the expected course grade. Indeed, the residual accounted for 95% of the variation in expected course grade after the fixed effects were taken into account. This result suggests that differences among the students contribute far more to expected course grades than factors such as the teacher, course section, or year. The means of the expected grade by levels of the demographic variables appear in Table 9, and the slopes for the continuous variables appear in Table 12.

Table 12. Slope of Covariates

Variable	**Expected grade including *study time this course***	**Expected grade without *study time this course***
Study time this course	–0.119	(Omitted from this model)
Study time all courses	NS	NS
Study habits	NS	NS
Chem/physics	0.066	0.068
Overall GPA	0.339	0.341

Students who had higher overall GPAs expected to receive higher grades in this course than those who had lower overall GPAs. Students with more previous successful coursework in chemistry and physics also expected to have higher end-of-semester grades than those who had less experience in chemistry and physics. Male students expected to get higher grades in the course than did female students. Furthermore, students 30 years of age or older anticipated higher grades than did traditional students. In the original model, which includes *study time this course*, students who spent more time studying for this course expected a lower grade than those who spent less such time.

6. CONCLUSIONS AND DISCUSSION

6.1 Student Evaluations

1. Grade inflation has the strongest correlation with evaluation items. Higher expected course grades relative to reported GPA result in more positive evaluations. (Remember that both *expected course grades* and *estimated GPA* are self-reported student items.)

The use of the *grade inflation* variable rather than *expected course grade* as an independent variable here makes it possible to reveal effects of grade inflation on evaluations. Assigning Bs to A students will produce very different reactions than assigning Bs to C students. This result is consistent with the results of Greenwald and Gillmore (1997a). (Some would argue that better instruction produces increased learning, which is reflected in higher evaluations.)

2. Stronger backgrounds in physical sciences negatively correlated with evaluation items. Students with more experience and success in past physical science courses submitted less positive evaluations.

   It is interesting that students with stronger backgrounds in the physical sciences rated the course lower than students with weaker backgrounds. Additional questions regarding the sources of dissatisfaction are being added to provide more insight here. It is quite likely that better-prepared students were not challenged by the pace of the course. Although an algebra-based version of this course, Astro 120, was added to the curriculum in 1996, many very talented students still enroll in Astro 100 either to minimize their effort or because Astro 100 offers more flexibility in scheduling (nine sections) as compared with Astro 120 (one section).

3. *Student study time* in this course is negatively correlated with evaluation items. Students who professed to spend more time studying for this course submitted less positive evaluations for both *course content* and *instructor mean* ratings.

   The negative correlation between student *study time for this course* and *course/instructor evaluation* items is almost surely linked to the fact that students who claimed to study more for the course received systematically lower grades. Potential explanations include inflation of reported study time and ineffective study habits. It is not uncommon for students to attempt to memorize course content rather than to develop generalized understanding of patterns and rules. Another possibility is that students do not like courses for which they are required to do substantial work outside of class.

4. There appear to be no statistically significant differences in evaluations that are related to sex or race.

6.2 Course Performance

1. Estimated GPA is positively correlated with expected course grade. (Remember that both *expected course grades* and *estimated GPA* are self-reported student items.)

   Estimated GPA is used as a measure of general academic ability in combination with effective mastery of academic skills. It is reassuring that this correlates strongly with expected course grades.

   Expected course grade is positively correlated with successful completion of previous chemistry and physics courses. It is logical that prior success in physical science courses correlates with course grade.

2. *Expected course grade* is negatively correlated with *study time for this course*. Students who professed to spend more time studying for this course expected to receive lower grades in this course.

   This was one of the more surprising results of the study. Because it seemed unlikely that increasing student study time *caused* students to get lower grades, the course performance model was repeated without *study time this course* as one of the independent variables. The results of this model are

provided in the tables above for comparison with the original model. The most conspicuous differences are the *F* factors for *race*, *age*, and *study time all courses* (Table 9). Differences related to race were more pronounced, differences related to age were less pronounced, and the effect of overall study time on the expected grade became virtually insignificant when *study time this course* was removed from the model. In this final model, the only significant factors were *GPA*, *prior success in chemistry and physics*, and *sex*.

The consistent negative correlation of study time for the course with course grade was unexpected but is consistent with results reported by Greenwald and Gillmore (1997b) and by Olivares (2002). If one is interested in interpreting the factors in the course performance model as causal agents that contribute to success, this negative correlation presents a real problem.

This result may reflect real differences related to the very broad range of abilities of the students enrolled in this course. High-quality students with strong backgrounds in the physical sciences can complete the course requirements without much effort. Students with lesser abilities and with weaker backgrounds in the physical sciences can spend significant time studying without significant improvement in exam scores.

This result might be explained, in part, if poor students inflated estimates of their study time for this course. It should be noted that in virtually all courses in the department, students estimated that their study time for this course constituted a disproportionately large share of their total study time. The student self-reported estimates of their study time are obviously very uncertain.

3. Female students reported lower expected course grades than male students.

   Statistically significant differences exist for gender and the two extreme age categories. In the case of the differences in expected course grade for male and female students, the results are consistent with calculations using final course grades for male and female students in selected sections that I (T. Robertson) taught. These calculations indicate that female students completed homework and unannounced quizzes at a higher rate than male students but received lower course grades of about 0.2 on a four-point scale.

The results of this study clearly demonstrate a positive correlation between student evaluation means and grade inflation, a correlation that should be recognized by those using student evaluation data for personnel decisions that include grant tenure and promotion, and awarding merit pay. It also reveals performance differences between the sexes, which merits more detailed study. The negative correlations between expected course grades and study time, and between student evaluation means and previous performance in chemistry and physics should lead to more carefully constructed studies that could result in pedagogical and curriculum changes.

The major strength of this assessment program has been in the development of data means for groups of similar courses within which individual instructors, department administrators, and promotion and tenure committee members can better evaluate the performance of faculty members.

Acknowledgments

This work was supported in part by funding from the Office of Academic Assessment at Ball State University. We would like to thank the referees and editorial staff for extensive suggestions that have significantly improved the article.

References

Bryk, A. S., & Raudenbush, S. W. 1992, *Hierarchical Linear Models: Applications and Data Analysis Methods*, Newbury Park, CA: Sage Publications.

Cashin, W. E. 1988, Student Ratings of Teaching: A Summary of the Research, Kansas State University Center for Faculty Evaluation and Development. IDEA Paper No. 20. http://www.idea.ksu.edu/index.html.

Cashin, W. E. 1995, Student Ratings of Teaching: The Research Revisited, IDEA Paper No. 32. Kansas State University Center for Faculty Evaluation and Development. http://www.idea.ksu.edu/index.html.

Cassady, J. C. 2001, Self-Reported GPA and SAT: A Methodological Note, *Practical Assessment, Research and Evaluation*, 7(12), http://PAREonline.net/getvn.asp?v=7&n=12.

Greenwald, A. G., & Gillmore, G. M. 1997a. 1997, Grading Leniency Is a Removable Contaminant of Student Ratings, *American Psychologist*, 52, 1209.

Greenwald, A. G., & Gillmore, G. M. 1997b. 1997, No Pain, No Gain?: The Importance of Measuring Course Workload in Student Ratings of Instruction, *Journal of Educational Psychology*, 89, 743.

Grodksy, E., & Gamoran, A. 2003, The Relationship between Professional Development and Professional Community in American Schools, *School Effectiveness and School Improvement*, 14(1), 1.

Gustafson, J. P., & Hendel, D. D. 2004, Using the Bryk and Raudenbush Model to Identify Effective and Ineffective Elementary Schools, Paper presented at the Annual Meeting of the American Educational Research Association meeting, San Diego, CA, April 12, 2004.

Hake, R. R. 2002, Problems with Student Evaluations: Is Assessment the Remedy?, http://listserv.nd.edu/cgi-bin/wa?A2=ind0204&L=pod&P=R14535.

Olivares, O. J. 2002, An Analysis of the Study Time-Grade Association, *Radical Pedagogy*, 4. http://radicalpedagogy.icaap.org/content/issue4_1/06_Olivares.html.