INT. J. SCI. EDUC., DECEMBER 2003, VOL. 25, NO. 12, 1509–1528

Taylor & Francis
Taylor & Francis Group

RESEARCH REPORT

Quantitative analysis of the rubric as an assessment tool: an empirical study of student peer-group rating

John C. Hafner, Moore Laboratory of Zoology and Department of Biology, Occidental College, Los Angeles, CA 90041, USA; e-mail: hafner@oxy.edu, and Patti M. Hafner, San Rafael Elementary School, Pasadena Unified School District, Pasadena, CA 91105, USA

Although the rubric has emerged as one of the most popular assessment tools in progressive educational programs, there is an unfortunate dearth of information in the literature quantifying the actual effectiveness of the rubric as an assessment tool *in the hands of the students*. This study focuses on the validity and reliability of the rubric as an assessment tool for student peer-group evaluation in an effort to further explore the use and effectiveness of the rubric. A total of 1577 peer-group ratings using a rubric for an oral presentation was used in this 3-year study involving 107 college biology students. A quantitative analysis of the rubric used in this study shows that it is used consistently by both students and the instructor across the study years. Moreover, the rubric appears to be 'gender neutral' and the students' academic strength has no significant bearing on the way that they employ the rubric. A significant, one-to-one relationship (slope = 1.0) between the instructor's assessment and the students' rating is seen across all years using the rubric. A generalizability study yields estimates of inter-rater reliability of moderate values across all years and allows for the estimation of variance components. Taken together, these data indicate that the general form and evaluative criteria of the rubric are clear and that the rubric is a useful assessment tool for peer-group (and self-) assessment by students. To our knowledge, these data provide the first statistical documentation of the validity and reliability of the rubric for student peer-group assessment.

Introduction and background for study

It seems that every academic discipline has its own, special lexicon, but in the jargon-rich field of education there is perhaps no word or phrase more confusing than the term 'rubric'. In the educational literature and among the teaching and learning practitioners, the word 'rubric' is understood generally to connote a simple assessment tool that describes levels of performance on a particular task and is used to assess outcomes in a variety of performance-based contexts from kindergarten through college (K-16) education (for example, Routman 1991, Herman et al. 1992, Pate et al. 1993, Custer 1996, Popham 1997, Finson and Ormsbee 1998). Although there are observed differences of opinion in the rubric's definition between the practitioners and the educational literature (for a review, see Wenzlaff, Fager and Coleman 1999), such differences pale in comparison with that found among college and university educators. Many college-level instructors working outside the discipline of education are often confused by the misappropriated use of the word 'rubric' by their colleagues in the education department and in the K-12 teaching community. It appears that the word 'rubric' has little (if any) pedagogical

International Journal of Science Education ISSN 0950–0963 print/ISSN 1464–5289 online © 2003 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/0950069022000038268

sex or success in the course affect how she or he assesses performance using the rubric? The intent of this contribution is to address these questions and others in an overall attempt to evaluate the effectiveness of the rubric when used by student raters in the actual setting of the biology classroom.

The study: methodology

The rubric used in this study (figure 1) was designed as both a standard-bearer and an assessment tool for collaborative oral presentations in the course 'Evolutionary Biology', a required course for Biology majors at Occidental College. Once constructed (see later), the rubric was used without modification for the past 3 years of this course (1998, 1999, and 2000), with class sizes of 36, 32, and 39 students, respectively. As a performance task, the students' oral presentations were considered by the instructor (J.C.H.) and the students as the culmination of their term-long collaborative research projects; the oral presentations were scheduled in the last 2 weeks of class and, together with their written research reports (due at the time of their oral presentation), the students' research projects were worth 150 of 500 total points in the course (oral presentation and written research report valued at 75 points each). Hence, the rubric was used in a situation where 'the constructed response being judged is fairly significant' (Popham 1997: 72).

The performance task

Students were asked to work cooperatively with another member of the class in small groups (typically pairs of students) to complete a library research project on a topic of shared interest in the realm of human evolutionary biology. Students were free to select virtually any topic in the broad realm of human evolutionary biology, but approval of the students' written 'Research Proposal' by the instructor was required prior to commencement of research to ensure appropriate scientific content. Details pertaining to the collaborative research project (as well as the rubric) were described in the course manual and discussed in class on several occasions. The research project involved library research, critical evaluation of published research in the primary literature, a written 'Interim Research Report', and culminated in oral and written presentations. For purposes of this study, it was the oral portion of the research project that represented the performance task and the focus of the rubric (figure 1).

Construction of the rubric

The rubric shown in figure 1 was modified from an earlier rubric that was constructed originally as part of a general education (Core) science course developed by the authors in 1997, entitled 'Introduction to Human Evolutionary Biology'. It is important to note that the rubric, in its original form (1997; not shown) and its modified form (1998; see figure 1), was constructed as a class activity with discussion and input from the students. Indeed, the five elements of the rubric, together with the evaluative criteria and the grading (point) system shown in figure 1, represented a consensus of student opinion; the instructor facilitated the discussion and provided the general grid design and total points (75) for the framework of the rubric. It should be noted that 'content' did not appear as a

Performance task—Students will work cooperatively to create an oral presentation on a topic of interest in the realm of human evolutionary biology. Although the content of each oral presentation will, of course, vary from presentation to presentation, the oral reports must follow the general guidelines discussed in class and presented in the Course Manual.

RUBRIC FOR AN ORAL PRESENTATION

ELEMENTS OF TASK	POINTS				
	15	13	11	9	7
Organization and research	Ideas are presented with clear sequence and logic; wide range of references; effective use of time	Organization is logical and sequential; adequate use of evidence; time budget is satisfactory	Organization displays lapses in logic and sequence; minimal use of resources; time management is wanting	Information is largely unorganized; very limited use of the literature; poor time management	Unorganized information without regard to logic or sequence; a dearth or lack of resources; time management absent
Persuasiveness and logic of argument	Persuasive stand; nice support and defense of thesis; opposing views are covered thoroughly	Clear thesis; covers topic thoroughly and with support; opposing views acknowledged	Thesis present; little support and defense of stand; opposing views not covered fully	Unreasonable or unclear thesis; wanders from topic	Rambles; unreasonable, unclear or missing thesis; no defense of stand
Collaboration	All members contributed eagerly to product; total participation; fulfilled assignment roles in the group	Adequate participation by all members; roles in group somewhat unclear; cooperation seems marginal	Off task at times; weak cooperation; one member with no or limited involvement	Dominated by one member; members seem not to have assigned roles in the group	Limited or no cooperation and participation; limited involvement in group process
Delivery and grammar	Enunciates clearly; no grammatical errors; well modulated; relaxed gestures; good eye contact; use of visuals	Enunciates clearly; diction is good; only rare errors in grammar; vocal modulation is fair to good	Enunciates clearly but simple sentences; fair vocal modulation; some mispronounced words; some errors in grammar	Stilted or slurred speech; modulation fair to poor; many words unclear; poor diction and grammar	Mumbles; modulation is poor (soft speech); generally cannot be understood; highly flawed grammar
Creativity and originality	Product demonstrates unusual insight and creativity; thoughtful and unique approach to topic	Product shows insight and creativity	Information presented is accurate with some evidence of creativity	Information is given with only very minimal creativity; product is basic with little elaboration	Product is very basic, displaying no creativity; no evidence of originality

Figure 1. The rubric for a collaborative oval presentation (the performance task) used in this study.

separate element in the rubric because of the highly varied nature of the scientific content of the presentations and because, by the time of the presentations at the end of the term, all projects had already been screened twice by the instructor for basic scientific content, first with the 'Research Proposal' and later with the 'Interim Research Report'. Nevertheless, scientific content was recognized as an important aspect of any presentation, and general descriptive criteria regarding content (e.g. logic, supporting references, coverage of topic and opposing views) were embedded in the first two elements of the rubric. After collaborative modification of the rubric in 1998 for the Evolutionary Biology course, the rubric was not further modified but students in subsequent classes (i.e. 1999 and 2000) were told of the history of the development and construction of the rubric both in the Course Manual as well as in class discussions. Although the authors do not claim that the rubric (figure 1) was flawless, it was judged to be useful as an instructional tool and seems to avoid the problems that plague many teacher-made and commercially published rubrics (for a discussion, see Popham 1997).

Use of the rubric

Students were introduced to the rubric early in the course and urged to consult the rubric and view it as a helpful guide when thinking about their oral presentations and working on their research projects. Each year, a mini-symposium entitled 'Human Evolutionary Biology' was held in the last 2 weeks of the course and it was here that the students presented their oral presentations. The mini-symposium was modeled after scientific conferences; a detailed schedule for the mini-symposium was issued well in advance of the date of the first presentation, all presentations were held in a fully equipped lecture room, and the instructor chaired all sessions.

Prior to the beginning of the symposium, the students were told that the instructor would greatly appreciate knowing their honest, confidential, peer-group assessment of the presentations using the rubric and that their assessments would be considered carefully and evaluated fully by the instructor (students were also told that the instructor would like to see their candid, self-assessment scores but that these scores would be excluded in final score determination). Upon entry to the lecture room for each day of the mini-symposium, each student was issued a one-page rating handout for the talks scheduled for that day. The students placed their names on this rating sheet and the students were asked to provide peer-group assessment of the student research presentations (up to four talks) for each day using the rubric. As a quick aid in scoring, a photographically reduced version of the rubric was reproduced at the top of the score sheet and students were asked to provide the presentation number (provided in the schedule and announced by the instructor while introducing each talk), the score on each of the five elements of the task, and the total points for the oral presentation. During the symposium, the instructor also rated the oral presentations using the rubric; hence, rating by the instructor was performed concurrently but independent of the peer-group rating by the students. At the end of each day's symposium, the students were asked to place their peer-group rating sheets face down on a shelf as they exited the lecture room and the score sheets were collected for later analysis (following conclusion of the course and assignment of grades). Beyond these simple instructions, the student raters were not given any training nor further directions in the use of the rubric.

Statistical analyses and handling of the data

Individual score sheets completed by the students for the peer-group assessment of student oral presentations and the instructor's rubric-based scores (INSTRUCTOR SCORE) for the same oral presentations provided the main quantitative basis for this analysis. A total of 52 oral presentations involving 107 students was evaluated by the use of the rubric. Data from the students' score sheets from each day of the mini-symposia were entered into the computer together with the students' sex (SEX), total points earned in the course (TOTAL POINTS, a measure of overall achievement), and year of class (1998, 1999, 2000); students' self assessments were omitted from these analyses. From the raw data, mean rating of talks by each student (STUDENT MEAN RATING) was calculated for all students, and these values provided a measure of grading rigor of the individual peer rater. Additionally, the mean peer scores (MEAN PEER SCORE) and other general descriptive statistics were calculated for each presentation.

Variables used in the analyses were examined for normality by use of the Lilliefors modification of the Kolmogorov–Smirnov test. For each year of the study and for the data pooled across years, TOTAL POINTS was found to depart significantly from the normal distribution; accordingly, all comparisons and analyses involving TOTAL POINTS were performed using non-parametric statistical routines (i.e. Kruskal–Wallace test, Mann–Whitney test, Spearman rank-order correlation, r_S). The variable STUDENT MEAN RATING was found to be distributed normally for individual years and for the pooled data except for year 1999.

To investigate the validity of the students' judgments of the oral presentations while using the rubric, the students' MEAN PEER SCORE for the presentations were compared with the INSTRUCTOR SCORE, the only other indicator of student oral performance available for the study. It should be emphasized here that the instructor's score, although employing the same rubric as used by the student raters, was independent of the students' scores and was the sole basis of grade determination on the oral presentation. Traditional regression analyses were employed when comparing MEAN PEER SCORE with INSTRUCTOR SCORE inasmuch as both variables were distributed normally for all sample years except for INSTRUCTOR SCORE for 1999. For the data pooled over the three years, both MEAN PEER SCORE and INSTRUCTOR SCORE were significantly non-normal in distribution; in this instance, non-parametric regression was performed using Kendall's robust line-fit method (Kendall and Gibbons 1990, Sokal and Rohlf 1995). In addition, Quenouille's ordering test was performed using the methodologically equivalent Kandall's rank correlation coefficient, r_K (Sokal and Rohlf 1995: 539).

Several methods were employed to evaluate inter-rater reliability, including conventional pair-wise correlation analyses using Spearman's r_S. The median r_S was used to summarize all pair-wise correlations among student raters per year because of the slightly skewed nature of the distributions. Friedman's two-way analysis of variance by ranks was also performed to test the null hypothesis of no systematic ranking of the oral presentations by the individual student raters. Because the Friedman's test requires a complete data set, mean values for a talk were substituted for appropriate missing values (e.g. all self-assessment scores and scores unavailable due to student absence). In addition, Kendall's coefficient of concordance, W, was used as a measure of overall correlation among the student raters.

relevance to the overwhelming majority of college-level and university-level teachers because of their academic appointment outside the education department and, in many cases, minimal preparation in teacher education; as a consequence, most professors in academia are usually unfamiliar with the popular pedagogical trends in 'alternative' and 'authentic' assessment of the past decade. To most college and university professors outside the realm of teacher education (together with the vast majority of the English speakers world wide), the term 'rubric' is understood most commonly to mean simply a title or major section of a book or manuscript. Although the term 'rubric' may not be the most appropriate descriptor for this assessment tool used by modern educators, the word is now embedded deeply in the vocabulary of K-12 teachers and teacher educators and, more importantly, the rubric is now emerging as a valuable pedagogical tool with which educators at all levels ought to be familiar.

With the obfuscating nature of the word aside, the rubric remains a popular topic in the educational literature, at educational conferences, and among K-12 teachers and teacher educators. Indeed, a significant body of literature has accumulated in the past decade on the design, construction, rationale, and use of the rubric as an alternative tool for assessment of performance (for example, Routman 1991, Wiggins 1991, Herman et al. 1992, Pate et al. 1993, Custer 1996, Popham 1997, Luft 1997, Stuhlmann et al. 1999). A rubric provides a description of various quantitative levels of performance for a performance task and describes what mastery (and varying degrees of mastery) of a performance task should look like (see Custer 1996, Luft 1997, Popham, 1997, Finson and Ormsbee 1998). The rubric is valuable to both the instructor and the student as a quick and clear summary of performance levels across a scoring scale. Importantly, the top level of the rubric communicates what exemplary work should look like and, as such, involves the student in constructive learning and self-evaluation.

A review of the literature shows, not surprisingly, that most quantitative analyses of the rubric focus on its design, concurrent validity, inter-rater reliability, and generalizability as a performance-assessment tool *from the teachers' perspective* (for example, Baker et al. 1995, Abedi and Baker 1995, Novak et al. 1996, Stuhlmann et al. 1999, Johnson et al. 2000). Such analyses are often 'laboratory studies' involving experienced raters, and designed mainly to evaluate and/or compare rubrics with the twin goals of enhancing teachers' instruction methods while, simultaneously, producing an assessment tool with technically defensible results.

Unfortunately, there is dearth of information in the literature on the actual effectiveness of the rubric in assessing performance when employed by student raters 'in the field'. Our interest here departs from previous studies and focuses on the validity and reliability of the rubric as a performance assessment tool *in the hands of the students*, and not on its use by teachers or other professional educators. Student peer-group raters, unlike teachers (whether trained or untrained in rubric use), may be considered as pedagogically naive raters. In the hands of these naive raters, is the rubric a valid and reliable tool for peer-group assessment of performance? Although the rubric is generally considered as an assessment tool for the instructor, we wish to examine the students' ability to use the rubric to assess different levels of performance among their peers. For example, to what extent do the students and the instructor agree on the assessments made for individual performances? Also, does a student's

Finally, a generalizability study for the presentation scores was conducted to assess student inter-rater reliability as well as to examine variance components. Such an analysis was carried out using conventional techniques for a two-way analysis of variance (ANOVA) without replication but followed the generalizability methods of Cronbach, Gleser, Nanda and Rajaratnam (1972), Shavelson and Webb (1991), and Brennan (1992). In the parlance of generalizability theory, we performed a single-facet, crossed $p \times r$ design (with rater being a 'facet') that allowed us to estimate variance components due to presentation (p) and rater (r), as well as error variance (residual variance due to interaction variance between presentation and rater, among other unknown factors). Additionally, generalizability analysis allowed for the calculation of two coefficients: the generalizability coefficient, ρ^2 (also termed the G-coefficient), which is analogous to the reliability coefficient from classical theory, and the index of dependability, ϕ.

Results

Profile of class

The Evolutionary Biology class for each year of this study represents an ethnically diverse sample of college students, principally sophomores and juniors majoring in biology. Mean class size over the three study years of 1998, 1999, and 2000 is 35.68 students, and individual class sizes are 36, 32, and 39 students, respectively. Overall, women outnumber men by roughly a two-to-one margin in the study (mean sex ratio = 2.06:1; a total of 72 females and 35 males) and sex ratios in individual classes all favor women (range 1.57–2.56:1).

Temporal variation in the class's achievement and rating behavior and the instructor's rating behavior is remarkably uniform. Indeed, the means for TOTAL POINTS earned by the students in the course for the years 1998, 1999, and 2000 are 424.722, 434.906, and 431.308, respectively, and there is no significant heterogeneity among these values (Kruskal–Wallis test statistic = 1.437, p = 0.487). For the same years, the means for MEAN PEER SCORE for students' oral presentations are 68.383, 69.252, and 69.460, respectively, and these means are not significantly different (Kruskal–Wallis test statistic = 1.522, p = 0.467). The students in each class also seem to use the rubric to evaluate their peers in a similar fashion across the study years as may be seen by examination of the means for STUDENT MEAN RATING. The mean values for STUDENT MEAN RATING for the years 1998, 1999, and 2000 are 68.294, 69.268, and 69.358, respectively, and there is no significant difference among these means (Kruskal–Wallis test statistic = 5.300, p = 0.071). Similarly, the instructor also seems to use the rubric consistently; from 1998 through 2000, the means for INSTRUCTOR SCORE for the presentations are 69.556, 70.375, and 68.556, respectively, and there is no significant heterogeneity among these means (Kruskal–Wallis test statistic = 1.387, p = 0.500).

Student participation and attendance at the mini-symposia across the three years of the study is strong. The total number of peer-group ratings using the rubric is 1577 and represents 89.86% of the total possible peer-group ratings assuming perfect attendance and subtracting self-assessments. Hence, absenteeism for the study was approximately 10%. This also represents a mean of 30.33 peer-group ratings (range = 23–37) for each of the 52 total talks evaluated in this study. It is

also interesting to note that the missing values (ratings by students) in the data set (due only to our culling of the students' self-assessments and absenteeism) are found to be missing completely at random (Little's [1988] MCAR test statistics [and probabilities] for years 1998, 1999, and 2000 are 381.919 [$p = 0.535$], 283.218 [$p = 0.502$], and 409.139 [$p = 0.419$], respectively).

Student achievement and leniency in peer assessment

To evaluate whether student overall performance in the course is associated with the relative leniency in rating of a student's peers, TOTAL POINTS is compared with STUDENT MEAN RATING for individual years as well as for pooled years. Analysis shows that there is no significant association between TOTAL POINTS and STUDENT MEAN RATING for individual years (1998, $r_S = -0.006$, $p \gg 0.05$; 1999, $r_S = 0.003$, $p \gg 0.05$; 2000, $r_S = -0.035$, $p \gg 0.05$) nor for pooled years ($r_S = 0.02$, $p \gg 0.05$).

Further analyses examining possible sex-based differences with regard to student achievement show that mean values for TOTAL POINTS for men and women for the individual years and pooled years show slightly greater means for women than for men students, but such values are not significantly different for any comparison except in one year (1998). In 1998, mean TOTAL POINTS for women (434.864 points) is significantly different from the mean for men (408.786 points) (Mann–Whitney $U = 221.50$, $p = 0.028$). An examination of possible differences between the sexes in their relative leniency in grading shows that there are no significant differences between the sexes for mean values for STUDENT MEAN RATING in any single year nor for the pooled data (for example, overall mean STUDENT MEAN RATING by males is 68.882 points [$n = 35$] and that by females is 69.017 points [$n = 72$]; $F_s = 0.088$, $p = 0.767$).

Validity

The validity of the rubric in the hands of the peer-group student raters is assessed with a single criterion (indicator): the instructor's independent and concurrent rubric scores, the sole basis of score (grade) assignment for the oral presentations. Specifically, regression analyses of the instructor's rating (INSTRUCTOR SCORE) on the students' peer-group ratings (MEAN PEER SCORE) for the presentations provides a check on the accuracy with which the students employ the rubric, as well as an overall assessment of the level of agreement between the students' and instructor's use of the rating tool. Year-by-year regression analyses (figure 2) show significant positive functional relations between INSTRUCTOR SCORE and MEAN PEER SCORE in all years: 1998 ($n = 18$ presentations), $b = 1.211$, $F_s = 57.494$, $p < 0.001$; 1999 ($n = 16$ presentations), $b = 1.143$, $F_s = 14.688$, $p = 0.002$; 2000 ($n = 18$ presentations), $b = 1.244$, $F_s = 10.082$, $p = 0.006$. Interestingly, each of the three regressions of INSTRUCTOR SCORE on MEAN PEER SCORE has a regression coefficient, b (slope), that is not significantly different from a parametric value of $b = 1.0$ (1998, $t_s = 1.32$, $p > 0.05$; 1999, $t_s = 0.48$, $p > 0.05$; 2000, $t_s = 0.62$, $p > 0.05$). Moreover, no heterogeneity is found among the three regression coefficients ($F_s = 0.023$, $p \gg 0.05$). Hence, all three slopes are considered to be sampled from the same statistical population. When the data are pooled across the study years, non-parametric regression of INSTRUCTOR

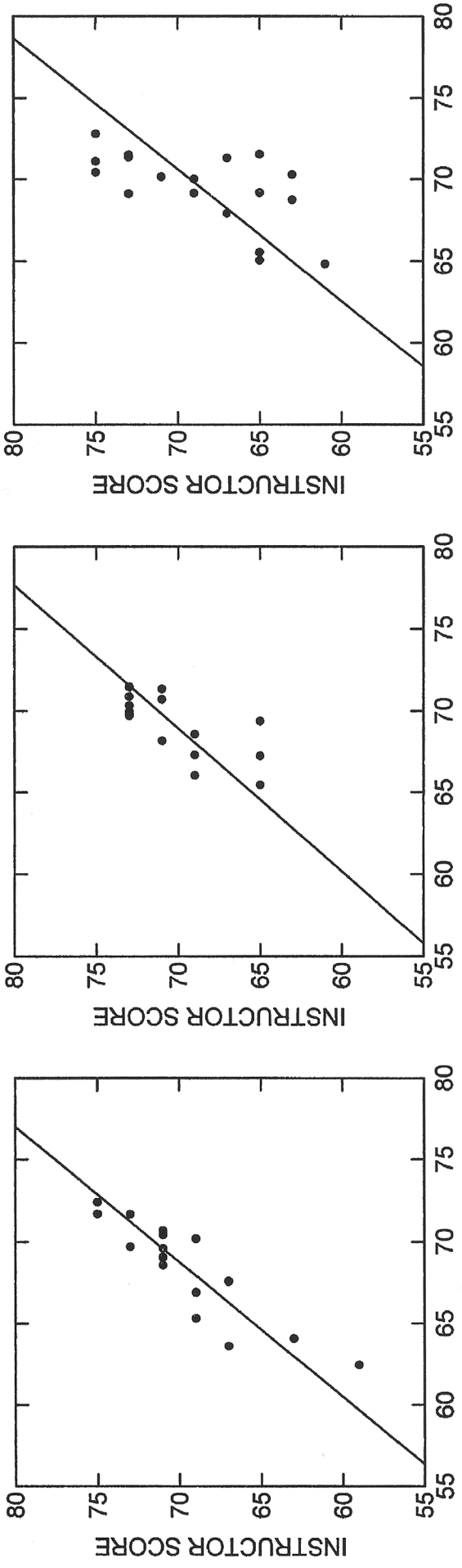

Figure 2. Linear regression analyses of INSTRUCTOR SCORE on MEAN PEER SCORE for the three years of the study: 1998, Y = 1.211X –13.263, r = 0.084; 1999, Y = 1.143X –8.765, r = 0.716; 2000, Y = 1.244X –17.850, r = 0.622. Regression analyses show virually a 1:1 relationship between the instructor's and the students' rating using the rubric (see text for discussion).

Table 1. Results of the generalizability study for the oral presentation scores and inter-rater reliability estimates for scores using the rubric. To facilitate comparison with Spearman's r_S, inter-rater reliability estimates from the generalizability study, ρ^2 and ϕ, are composed of scores from two raters.

	Variance components (percent)			*Generalizability coefficients*		
Year	*Presentation*	*Rater*	*Error*	ρ^2	ϕ	*Median* r_S
1998	7.109 (36.5)	2.586 (13.3)	9.765 (50.2)	0.5928	0.5351	0.500
1999	2.892 (18.8)	4.560 (29.6)	7.954 (51.6)	0.4210	0.3161	0.366
2000	4.573 (25.7)	3.676 (20.7)	9.516 (53.6)	0.4901	0.4094	0.387

SCORE on MEAN PEER SCORE yields a significant slope of similar magnitude and sign ($b = 1.166, r_K = 0.548, p \ll 0.01$).

Inter-rater reliability

The Friedman's two-way analysis of variance by ranks for each year of the study shows that students agree strongly in their ability to assess different levels of performance in the oral presentations using the rubric. For years 1998, 1999, and 2000, a true preferential order (systematic ranking) of the symposium's oral presentations is identified by the students as is documented by highly significant test statistics (Friedman test statistics for the three years are 303.614 [$p < 0.001$], 166.091 [$p < 0.001$], and 250.471 [$p < 0.001$], respectively). Hence, the students show much agreement in their ranking of the oral presentations and, in all cases, the null hypothesis of no systematic ranking is rejected. Significant inter-rater reliability is also demonstrated by Kendall's coefficient of concordance, *W*; Kendall's *W* for the years 1998–2000 is 0.496, 0.346, and 0.378, respectively ($p \ll 0.01$ in all cases). Finally, correlations between rater pairs, a more traditional measure of inter-rater reliability, provides further evidence of agreement in rater scores. For 1998, the median correlation, r_S, across raters is 0.500, and is 0.366 and 0.387 for the following two years, respectively.

Results of the generalizability study are presented in table 1. The generalizability analysis allows for the dissection of total variance in scores into variance components; total variance is partitioned into variance among presentation, rater variance, and error variance. The general pattern of variance components across the three years shows that approximately one-quarter (range, 18.8–36.5%) of the total variance is due to differences among the presentations, another one-quarter (range, 13.3–29.6%) of the total variance is accounted for by rater differences, and approximately one-half (range, 50.2–53.6%) of the total variance is residual error (see table 1). The partitioning of variance through the generalizability study also allows for the computation of generalizability coefficients, ρ^2 and ϕ. Table 1 also provides these generalizability coefficients for each year of the study, along with their corresponding correlation coefficients (median r_S), as further corroboration of inter-rater reliability.

Discussion

Conversations with students over the past 3 years (as well as examination of students' written evaluation of the course) reveal that the students readily accept and use the rubric. Indeed, it appears that the students find the rubric to be a very helpful study tool. Students appreciate knowing what is expected of them and how they are to be evaluated for their performance. Although comparative data on the performance of the class without the rubric are unavailable, it is our opinion that the overall quality of the presentations is higher and attendance is better at the symposia because of the use of the rubric.

Possible biases in use of the rubric

Quantitative analysis of the rubric used in this study (figure 1) shows that it is a measurement tool that is used consistently by both the student and the instructor across the study years (no significant differences in STUDENT MEAN RATING nor INSTRUCTOR SCORE for the three years). Also, the students' overall academic strength in the course (as measured by TOTAL POINTS) has no significant bearing on the way that they employ the rubric (STUDENT MEAN RATING) to assess their fellow students. Moreover, the rubric appears to be 'gender neutral' inasmuch as there is no significant difference between the sexes in their use of the rubric as a scoring tool; hence, neither sex is more or less lenient than the other in peer rating. This latter finding is of interest because it is now documented thoroughly across a variety of assessment programs that men and women perform differently when responding to questions on constructed response and multiple-choice testing formats (for example, Breland and Griswold 1982, Breland et al. 1994, Pomplun and Capps 1999, Pomplun and Sundbye 1999). Given that men and women perform differently *when being evaluated* in conventional testing situations, it is of interest in the present context to know whether men and women perform differently *when they are evaluating others* using the rubric; again, our results show that they assess others with equal rigor.

Validity and reliability

In the realm of applied statistics, it is well recognized that the usefulness of a measurement is rooted both in its accuracy, the closeness of a single measure to the true value, and in its precision (often termed consistency, reliability, or comparability), the closeness of repeated measures of the same item to each other (for example, Simpson et al. 1960, Snedecor and Cochran 1989, Hopkins et al. 1990, Sokal and Rohlf 1995, Zar 1996). In education research and psychometrics, it appears that the terms 'accuracy' and 'precision' are generally eschewed in favor of the more specific terms of 'validity' and 'reliability', when it comes to examining the usefulness of a tool in measuring human psychological performance. Ebel (1972: 444) defines validity as 'the accuracy with which a set of test scores measures what it ought to measure' and reliability as 'the consistency with which a set of test scores measures whatever it does measure'. Despite terminological differences across the disciplines, it is clear to all practitioners that a quality measurement or assessment tool must be both valid (accurate) and reliable (precise) if it is to be useful.

Our goal here is to examine the validity and reliability of the rubric as an assessment tool for use by the students. Because of the design of the mini-symposia in the course, we had but a single criterion to assess validity of the peer-group scores using the rubric: the instructor's independent ratings of the oral presentations that were made using the same rubric and scored concurrently. Although grade-point average is often used as the criterion of choice to assess concurrent validity of a performance, we judge its use here to be inappropriate mainly because the grade-point average is overwhelmingly biased in favor of assessments of written and analytic expression, and not oral performance (for a discussion, see Abedi 1991, Abedi and Baker 1995). Given that the instructor's scores are independent of the students' scores (in the sense that 'blind scoring' was used) and reflect expert judgment (rooted both in familiarity with rubrics as well as nearly 20 years of college teaching experience), the instructor's ratings provide a means of assessing the validity of the students' peer-group scores using the rubric. With some latitude, this measure of validity may be viewed as 'concurrent validity', although this usage may depart from more restricted and traditional definitions (for example, Hopkins et al. 1990, Walsh and Betz 1990) in that the scores of the criterion (i.e. the instructor's score) are not based on a separate test or measure. Terminology aside, significant regression analyses demonstrate remarkable validity of the rubric in the hands of the students; students' scores are able to predict instructor's scores with accuracy (a 1:1 relationship is demonstrated) and with moderate ($r = 0.622$) to high ($r = 0.884$) correlation coefficients across the study years (see figure 2). The 1:1 relation is particularly noteworthy in our case (with students and instructor employing the same rubric) because it demonstrates that the students, although naive raters, employ the rubric in the same manner as does the instructor; hence, student peer-group rating using the rubric is a valid assessment of performance and, in turn, a strong predictor of their grade on this task.

Multiple indices support the reliability of the peer-group rubric scores. One measure of reliability is the level of agreement between pairs of student raters on the rank ordering of the presentations. As was done by Novak et al. (1996), correlations between rater pairs assess the stability of ranking of presentations across different raters and show median r_S coefficients of approximately 0.4–0.5 (see table 1). Admittedly, the magnitudes of these values do not suggest high reliability (as a value of $r_S \geq 0.9$ would indicate), but only moderate reliability. It is important to bear in mind that, although a higher correlation coefficient is always preferred, the student raters in this study are naive raters, no discrepant scores whatsoever were culled from the analyses, and this assessed performance (oral presentation) is but one aspect of a composite grade for the course (for a discussion, see Hopkins et al. 1990). Clearly, our reliability estimates for assessment of oral presentations are lower than one desires for, say, standardized tests but they are, for example, similar in magnitude to values obtained from portfolio assessment (approximately 0.40; Koretz et al. 1994) and writing assessment (0.45–0.69; Novak et al. 1996).

Friedman's tests and Kendall's W also show that the rubric, when used for peer evaluation, has high comparability. The Friedman's tests show that the students' rankings of the performances have a true preferential order; that is, students are able to recognize consistently higher and lower quality performances such that there is strong (highly significant) agreement in the ranking order of the presentations for each year. Kendall's coefficient of concordance, W, further affirms the reliability of

the scores with highly significant values of W of approximately 0.4–0.5 (again, moderate levels of concordance).

Estimates of inter-rater reliability from the generalizability study further support the stability of the rubric for peer-group assessment. As may be seen in table 1, the generalizability coefficient, ρ^2 and the dependability coefficient, ϕ, show general agreement with the median correlation coefficients, r_S, across all years. Interestingly, the first year of the study, 1998, shows higher values for the three inter-rater reliability estimates than the corresponding values for the other two years (table 1). One possible explanation for the higher reliability estimates for 1998 is that this class, unlike the others, participated in the actual construction of the rubric. The in-class exercise of building the rubric may have given the class of 1998 a heightened sense of ownership and a deeper understanding of the evaluative criteria and elements of the rubric than the other classes and may explain the higher reliability estimates for that year.

Generalizability and decision studies

The principal thrust of any generalizability study is to estimate variance components, whereas the purpose of a decision study is to evaluate those variance components for various decision-making purposes regarding assessment (see Cronbach et al. 1972, Shavelson and Webb 1991, Brennan 1992). The estimated variance components of table 1 provide useful information when generalizing from a presentation's score by a single rater to its universe score (true score) that would be obtained, in theory, if averaged over all raters in the universe. In our generalizability study, the variance component for presentation represents the 'true-score' or 'universe-score' variance and accounts for 18.8–36.5% of the total variance depending on the year (table 1); hence, we may hypothesize that roughly one-quarter of the total variance in scores on the presentations is due to true capability in performance with about three-quarters of the total variance in scores being attributable to other factors (i.e. differences in scoring among the raters and residual variance). The variance component for rater contributes an estimated 13.3–29.6% of the total variance across years or, again, approximately one-quarter of the total variance in scores. The relative magnitude of rater variance indicates that a substantial proportion of the total variance is due to varying leniency of raters; such rater differences, however, appear to be independent of sex and overall academic performance in the course (see earlier). The residual variance accounts for approximately one-half of the total variance across the years and is remarkably consistent (table 1). Residual variance includes all error variance that is due to any presentation–rater ($p \times r$) interaction and from other unknown sources not addressed in our one-facet study. A consistently large residual variance component across years suggests to us that there is a certain level of discrepancy in ranking of the presentations across the raters (but recall the significant results of the Friedman's test and Kendall's W) and that there may be other systematic factors (e.g. academic class [sophomore, junior, or senior] or number of previous biology courses taken) or unsystematic factors (e.g. random discrepant scores occasionally submitted by inattentive or distracted peer raters) that we cannot address due to the design of this study.

Further examination of the estimated variance component for presentation (universe) scores across the three years (table 1) reveals a large (36.5%) variance in

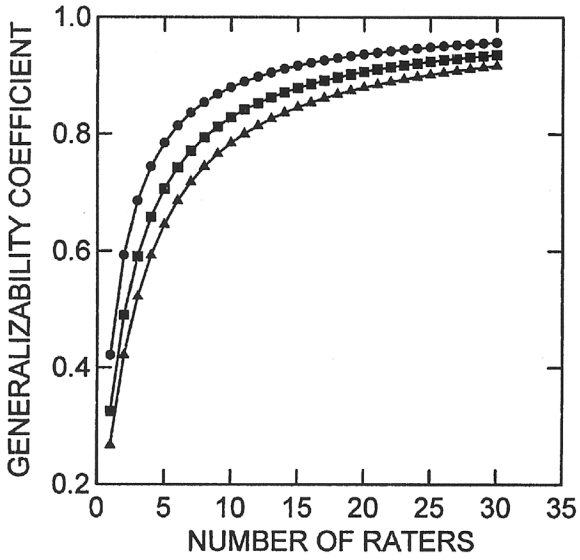

Figure 3. Score generalizability of the rubric as a function of number of raters (generalizability coefficients, ρ^2, are estimated from variance components). Data from the 1998, 1999, and 2000 classes are represented by circles, triangles, and squares, respectively. Note that as few as 5–10 raters yield moderately high generalizability coefficients for all three years.

the first year of the study (1998), with smaller values in succeeding years. Given that the 1998 class was the only class of the three that actually participated in the construction of the rubric, we interpret the high variance component for presentation in 1998 as a possible reflection of the students' higher level of understanding of the rubric than students in the following years (whom were told about rubric construction but did not participate directly). Although one cannot rule out chance (sampling error) as a factor in explaining the magnitude differences in presentation variance (only three study years available), it may be that the greater variance in presentation scores of 1998 is due to the peer raters being able to discern differing levels of performance more accurately than raters in other years (see figure 2; correlation between INSTRUCTOR SCORE and MEAN PEER SCORE for 1998 is $r = 0.884$).

As noted earlier, estimates of variance components from the generalizability study provide the basic 'raw material' for a decision study. Specifically, a decision study allows for the estimation of generalizability coefficients (ρ^2 and ϕ) as well as a consideration of the confidence that one may have when generalizing from a score by a single peer rater to the universe (true) score on a presentation. The estimates for the generalizability coefficients of table 1 are based on two raters in keeping with convention (see Johnson et al. 2000) and to aid in comparing these coefficients with the more traditional Spearman's correlation coefficient (r_S is limited to inter-rater reliability based on rater pairs). Again, these estimates for ρ^2 and ϕ across the three study years are of only mediocre value, they agree with the reliability measures for r_S, and express the estimated level of generalizability and dependability assuming two raters.

Information from the generalizability study may also be used in the context of a decision study to evaluate ways of improving the reliability of the peer-scored rubric. In the realm of decision studies, one may be interested in optimizing the

number of conditions of a facet (i.e. the number of peer raters) to attain a particular level of reliability. It is well known from classical test theory that the reliability of an assessment may be improved by relying on scores averaged over multiple raters. Importantly, generalizability theory allows for the estimation of generalizability coefficients that predict the reliability of scores as a function of the number of raters. Figure 3 shows the theoretical increase in reliability (ρ^2) of the rubric scores with increasing number of raters for the three years of study. Note that, in all cases, there is a rapid gain in score generalizability from a single rater to about five raters but that the incremental increase in reliability levels off with approximately 15 raters. In the present study, we can be confident that the values of MEAN PEER SCORE for the presentations show high generalizability (theoretically, $\rho^2 > 0.9$) inasmuch as the mean number of peer-group raters per presentation was approximately 30 raters. More importantly, this decision study shows that moderately high generalizability coefficients ($\rho^2 \approx 0.80$) are predicted with class sizes (peer-group raters) as small as only about 10 students (figure 3) when using this rubric. Such knowledge is invaluable when considering future designs for efficient assessment programs using this rubric. For example, peer-group assessment utilizing this rubric should provide moderately high generalizability when used in small, discussion classes, laboratories, and/or seminars with as few as 10–15 students.

Quality of performance and variance in peer rating

Although it is pleasing to see that the students are able to use the rubric for peer-group assessment with a high degree of accuracy and comparability, a close examination of the data reveals a somewhat surprising finding; there is a strong inverse relation between MEAN PEER SCORE and variance in peer score. Indeed, figure 4 shows that there is a significant decreasing trend between MEAN PEER

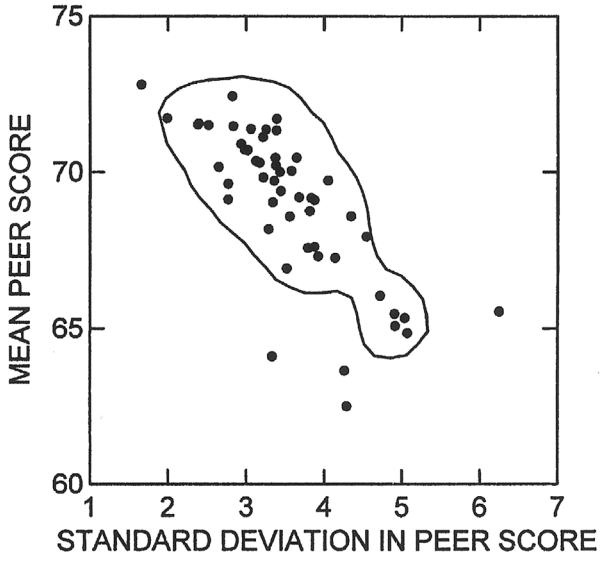

Figure 4. Inverse relationship between quality of performance (MEAN PEER SCORE) and variance in peer score ratings (as shown by STANDARD DEVIATION IN PEER SCORE) across the three years of the study (52 presentations). The contour line depicts the 75% confidence region for the data and is constructed from nonparametric kernel density estimators.

SCORE and standard deviation (= square root of variance) in peer score (r_K = −0.618, $p \ll 0.01$; r_S = −0.791, $p \ll 0.01$) for the study years. We note two possible explanations for this inverse relation between the quality of the oral presentations (as measured by MEAN PEER SCORE) and the variance in the ratings given by peer scorers. First, it may be easier for the students to recognize more superior performances than it is for them to recognize poorer quality presentations. Our reasoning here is that virtually all students may respond favorably and consistently to the higher quality talks but, because of the different levels of intellectual sophistication of the students, subtle conceptual errors of the less superior talks are detected by some students but pass unnoticed by other students. A second, and maybe more likely, explanation for this inverse relation is 'the sympathy factor'. Perhaps there is increase variance in rating of the lesser-quality presentations because some students opt to give slightly higher scores to these talks than they actually deserve because of some sort of socially constructed but inappropriate sense of 'kindness'. Whether either or both of these two explanations (or some unknown factors) are responsible for this inverse relation is interesting from a psychological perspective but beyond the scope of the present study.

Assessed performance and sequence of talk

Close analysis of the data shows another somewhat surprising finding. There is a significant positive association between the sequence of the oral presentation during the symposium and the assessed rating (both by peer-group assessment and by the instructor). Figure 5 shows the weak but significant positive association between MEAN PEER SCORE and sequence of talk of all 52 talks for the study (r_S = 0.396, $p < 0.01$), such that students giving their talks later in the sequence may

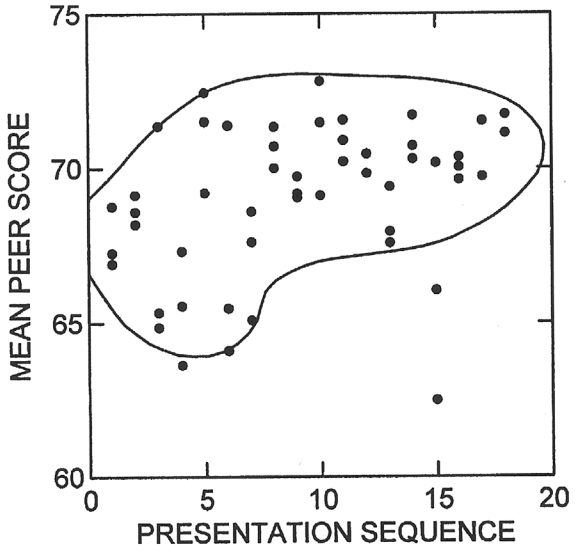

Figure 5. Positive association between quality of performance (MEAN PEER SCORE) and order of the presentation in the sequence of talks of the mini-symposia (PRESENTATION SEQUENCE). Contour line shows the 75% confidence region for the data and is constructed from nonparametric kernel density estimators.

have a slight advantage over those presenting earlier. This relation also holds for INSTRUCTOR SCORE and talk sequence across the three years but the relation is even more weak ($r_S = 0.292$, $p < 0.05$); the trend is generally not significant on a per-year basis.

This significant, positive relation between rating and talk sequence is somewhat troubling given all the information about the rubrics and the oral presentations provided in the course manual and discussed in class. Although the association is weak, students may perform slightly better later in the sequence of talks because of three possible reasons: (1) they may watch and learn from earlier performances; (2) they may have slightly more time to prepare; and (3) the students have 'trained' themselves to better use the rubric. While the factor of additional prep time cannot be addressed, that of visual demonstration and training can be considered in the design of future assessments of this sort. Because of the various learning modalities of students, the future inclusion of videotaped 'benchmark' talks could be advantageous to some individuals. These videos, coupled with the details outlined in the rubric, would serve to demonstrate various performance levels to the students and allow them to gain a consistent understanding of the rubric. To truly be effective, the videos should also have a narrative that explains how the rubric was applied in its various components. It is our thinking that these benchmark talks might be taped subsequent to their initial presentation and with the students' knowledge.

Student self-assessment

Although peer-group rating using the rubric is our primary focus, it is also interesting to examine the students' self-assessment scores using the rubric. For the three years of the study, 61 of 107 (57%) students submitted self-assessment rubric scores; four self-assessment rubrics were culled from analysis because they were judged to be bogus scores (maximum points awarded to themselves, coupled with silly comments), leaving a total of 57 self-assessment rubrics for examination. To assess validity, the self-assessment scores were compared with INSTRUCTOR SCORE and with TOTAL POINTS earned in the course. Given the non-normality of the self-assessment scores, non-parametric regression analyses were performed; Kendall's robust line-fit method, coupled with Quenouille's (1952) ordering test, shows a significant, positive slope of unity for the regression of INSTRUCTOR SCORE on the self-assessment scores ($b = 1.000$, $r_K = 0.250$, $p < 0.01$). Although the mean of the self-assessments scores (70.439) is slightly higher than that of the instructor's score (69.491), such means are not significantly different ($U = 1450.5$, $p = 0.315$) as would be expected with a slope of unity between the variables. Quenouille's ordering test also shows that there is a significant increasing trend between self-assessment scores and TOTAL POINTS ($r_K = 0.247$, $p < 0.01$). Taken together, these data indicate that use of the rubric for self-evaluation affords the students valid assessments of their own performances.

Some implications

Most college science courses, whether consciously or unconsciously, take a progressive, multifaceted approach to learning due to their lecture plus laboratory and/or field formats (for example, Hafner and Hafner 1992); this 'hands-on'

approach has served science well and is an approach that is now emulated in the humanities and recognized widely for its pedagogical virtues. Despite the many virtues of this approach, assessment in science courses, for the most part, consists of traditional examinations and term papers where the grade is determined by 'semi-secret devices (gradebooks, test booklets, etc.) containing information intended for the teacher's eyes' (Custer 1996: 30). Importantly, the inclusion of the rubric adds a different dimension to the assessment process by providing the student with a self-assessment tool before and during the construction of the project.

There is a surfeit of information in the educational literature about the virtues of the rubric. Unfortunately, the overwhelming majority of these studies and commentaries focus on *the teacher assessing* and not on *the student learning*. It is this latter approach, the use of the rubric from the student's perspective, that we have sought to explore. As educators, it is our job to help students understand how to construct their own learning and thereby continue to be life-long learners. The use of rubrics as a teaching and learning tool can play an integral part in attaining this goal.

Peer assessment may also play an integral role as students continue to develop life-long learning skills. Although an evaluation of the merits of peer assessment are beyond the scope of this contribution, the benefits of peer learning and peer assessment are far reaching and documented thoroughly in the literature. Peer learning and peer assessment are key pedagogical strategies to help students gain the knowledge that allows them to reflect on their own performance as well as that of their fellow students (Lejk and Wyvill 2001, Sluijsmans et al. 2001). Additionally, research has shown that group work and peer assessment have resulted in increased student comprehension and, ultimately, in higher grades (for a review, see Falchikov 1991, Gatfield 1999).

Our experience leads us to conclude that the use of the rubric in combination with peer assessment provides an effective teaching and learning strategy for this performance task (an oral presentation) in the setting of a college science classroom. Although the design of the present study does not provide empirical data pertaining to comprehension associated with the use of peer assessment and/or the rubric relative to situations without either or both approaches, it is our clear impression that these approaches do indeed foster scholarly presentations, increased audience participation, and an overall, enhanced level of student achievement. Future studies assessing quantitatively this hypothesized relationship should provide a promising avenue for further research.

Acknowledgments

This study was made possible by a curriculum development grant from the Los Angeles Collaborative for Teacher Excellence (LACTE). LACTE is a project funded by the National Science Foundation's program Collaboratives for Excellence in Teacher Preparation (CETP). We appreciate the support from the campus co-coordinators for LACTE, Don Goldberg, Laurie Fathe, and Mickey McDonald, and we are also grateful to the students of the Human Evolutionary Biology (Biology 104) class of Spring 1997 who served as the 'guinea pigs' for the initial development of the rubric used in this study. Dian Teigler provided much appreciated assistance in library research and information services. Finally, we would like to thank all the Evolutionary Biology (Biology 279) students in the

Spring 1998, Spring 1999, and Spring 2000 classes at Occidental College for their cooperation and support of the rubric; without their faithful and conscientious use of the rubric, this study would not have been possible.

References

ABEDI, J. (1991). Predicting graduate academic success from undergraduate academic performance: a canonical correlation study. *Educational and Psychological Measurement*, 51, 151–160.

ABEDI, J., and BAKER, E. L. (1995). A latent-variable modeling approach to assessing interrater reliability, topic generalizability, and validity of a content assessment scoring rubric. *Educational and Psychological Measurement*, 55, 701–715.

BAKER, E. L., ABEDI, J., LINN, R. L., and NIEMI, D. (1995). Dimensionality and generalizability of domain-independent performance assessments. *The Journal of Educational Research*, 89, 197–205.

BRELAND, H., and GRISWOLD, P. A. (1982). Use of a performance test as a criterion in a differential validity study. *Journal of Educational Psychology*, 74, 713–721.

BRELAND, H., DANOS, D., KAHN, H., KUBOTA, M., and BONNER, M. (1994). Performance versus objective testing and gender: an exploratory study of an Advanced Placement History examination. *Journal of Educational Measurement*, 31, 275–293.

BRENNAN, R. L. (1992). An NCME instructional module on generalizability theory. *Educational Measurement: Issues and Practice, Winter*, 27–34.

CRONBACH, L. J., GLESER, G. C., NANDA, H., and RAJARATNAM, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles* (New York: John Wiley and Sons).

CUSTER, R. L. (1996). Rubrics: an authentic assessment tool for technology education. *The Technology Teacher, December/January*, 27–37.

EBEL, R. L. (1972). *Essentials of Educational Measurement* (Englewood Cliffs, NJ: Prentice-Hall).

FALCHIKOV, N. (1991). Group process analysis: self and peer assessment of working together in a group. In S. Brown, and P. Dove (Eds.), *Self and peer assessment.* Presented at Birmingham, Standing Conference on Educational Development (Paper No. 63).

FINSON, K. D., and ORMSBEE, C. K. (1998). Rubrics and their use in inclusive science. *Intervention in School and Clinic*, 34, 79–88.

GATFIELD, T. (1999). Examining student satisfaction with group projects and peer assessment. *Assessment and Evaluation in Higher Education*, 24, 365–377.

HAFNER, J. C., and HAFNER, M. S. (1992). Laboratory investigations and discussions: an alternative pedagogical strategy in evolutionary biology. In R. G. Good, J. E. Trowbridge, S. S. Demastes, J. H. Wandersee, M. S. Hafner, and C. L. Cummins (Eds.), *Proceedings of the 1992 Evolution Education Research Conference* (Baton Rouge: Louisiana State University).

HERMAN, J. L., ASCHBACHER, P. R., and WINTERS, L. (1992). *A Practical Guide to Alternative Assessment* (Alexandria, VA: Association for Supervision and Curriculum Development).

HOPKINS, K. D., STANLEY, J. C., and HOPKINS, B. R. (1990). *Educational and Psychological Measurement and Evaluation* (7th ed.) (Englewood Cliffs, NJ: Prentice-Hall).

JOHNSON, R. L., PENNY, J., and GORDON, B. (2000). The relation between score resolution methods and interrater reliability: an empirical study of an analytic scoring rubric. *Applied Measurement in Education*, 13, 121–138.

KENDALL, M. G., and GIBBONS, J. D. (1990). *Rank Correlation Methods* (5th ed.) (London: Edward Arnold).

KORETZ, D., MCCAFFREY, D., KLEIN, S., BELL, R., and STECHER, B. (1994). *The Reliability of Vermont Portfolio Scores in the 1992–93 School Year*, Report No. TM021562 (Los Angeles, CA: The Rand Corporation, National Center for Research on Evaluation, Standards, and Student Testing).

LEJK, M., and WYVILL, M. (2001). Peer assessment of contributions to a group project: a comparison of holistic and category-based approaches. *Assessment and Evaluation in Higher Education*, 26, 61–72

LITTLE, R. J. A. (1988). Robust estimation of the mean and covariance matrix from data with missing values. *Applied Statistics*, 37, 23–28.

LUFT, J. (1997). Design your own rubric. *Science Scope*, 20, 25–27

NOVAK, J. R., HERMAN, J. L., and GEARHART, M. (1996). Establishing validity for performance-based assessments: an illustration for collections of student writing. *The Journal of Educational Research*, 89, 220–233.

PATE, P. E., HOMESTEAD, E., and MCGINNIS, K. (1993). Designing rubrics for authentic assessment. *Middle School Journal*, 25, 25–27.

POMPLUN, M., and CAPPS, L. (1999). Gender differences for constructed-response mathematics items. *Educational and Psychological Measurement*, 59, 597–614.

POMPLUN, M., and SUNDBYE, N. (1999). Gender differences in constructed response reading items. *Applied Measurement in Education*, 12, 95–109.

POPHAM, W. J. (1997). What's wrong–and what's right–with rubrics. *Educational Leadership, October*, 72–75.

ROUTMAN, R. (1991). *Invitations: Changing as Teachers and Learners K-12* (Portsmouth, NH: Heinemann).

SHAVELSON, R. J., and WEBB, N. M. (1991). *Generalizability Theory: A Primer* (Newbury Park, CA: Sage Publications).

SIMPSON, G. G., ROE, A., and LEWONTIN, R. C. (1960). *Quantitative Zoology* (rev. ed.) (New York: Harcourt, Brace and World).

SLUIJSMANS, D. M. A., MOERKERKE, G., VAN MERRIENBOER, J. J. G., and DOCHY, F. J. R. C. (2001). Peer assessment in problem based learning. *Studies in Educational Evaluation*, 27, 153–173.

SNEDECOR, G. W., and COCHRAN, W. G. (1989). *Statistical Methods* (8th ed.) (Ames, IA: Iowa State University Press).

SOKAL, R. R., and ROHLF, F. J. (1995). *Biometry* (3rd ed.) (New York: W. H. Freeman and Company).

STUHLMANN, J., DANIEL, C., DELLINGER, A., DENNY, R. K., and POWERS, T. (1999). A generalizability study of the effects of training on teachers' abilities to rate children's writing using a rubric. *Journal of Reading Psychology*, 20, 107–127.

WALSH, W. B., and BETZ, N. E. (1990). *Tests and Assessment* (2nd ed.) (Englewood Cliffs, NJ: Prentice-Hall).

WENZLAFF, T. L., FAGER, J. J., and COLEMAN, M. J. (1999). What is a rubric? Do practitioners and the literature agree? *Contemporary Education*, 70, 41–46.

WIGGINS, G. (1991). Standards, not standardization: evoking quality student work. *Educational Leadership, February*, 18–25.

ZAR, J. H. (1996). *Biostatistical Analysis* (3rd ed.) (Upper Saddle River, NJ: Prentice Hall).