<http://dx.doi.org/10.1080/15512160590961784>. Open Access Version: <http://kuscholarworks.ku.edu/dspace/>.

Self and Peer Evaluation in Undergraduate Education: Are Promises Worth Risking the Perils?

Mariya Y. Omelicheva

Paper citation:

Omelicheva, Mariya Y. Self- and Peer-Evaluation in Undergraduate Education: Structuring Conditions That Maximize Its Promises and Minimize the Perils, *Journal of Political Science Education* 1(1): 191-206, 2005.

Abstract:

This study canvasses reliability of students' self and peer evaluation, a method of assessment of university students that has recently gained renewed pedagogical interest and broad recognition. Two experiments, imbedded in classroom curriculum, examined the effects of the instrument of evaluation (with criteria vs. no criteria for evaluation provided), the format of evaluation procedure (anonymous vs. non-anonymous), and motivation of students (strong vs. weak) on the accuracy of students' self and peer ratings. The results of the experiments revealed both a considerable unreliability of peer ratings in some cases as well as a notable consistency of peer evaluations in others. The instrument of evaluation with criteria provided had significant positive effect on the accuracy of peer evaluations. This finding was robust across both experiments reported in the paper. Students' motivation also had impact on the reliability of peer evaluations. Students strongly motivated to apply criteria for evaluation produced more accurate peer evaluations compared to their peers provided with not criteria or supported with the criteria but not motivated to apply them. The results on the impact of the condition of anonymity were mixed.

This study canvasses reliability of students' self and peer evaluation, a method of assessment of university students that has recently gained renewed pedagogical interest and broad recognition. Two experiments, imbedded in classroom curriculum, examined the effects of the instrument of evaluation (with criteria vs. no criteria for evaluation provided), the format of evaluation procedure (anonymous vs. non-anonymous), and motivation of students (strong vs. weak) on the accuracy of students' self and peer ratings. The results of the experiments revealed both a considerable unreliability of peer ratings in some cases as well as a notable consistency of peer evaluations in others. The instrument of evaluation with criteria provided had significant positive effect on the accuracy of peer evaluations. This finding was robust across both experiments reported in the paper. Students' motivation also had impact on the reliability of peer evaluations. Students strongly motivated to apply criteria for evaluation produced more accurate peer evaluations compared to their peers provided with not criteria or supported with the criteria but not motivated to apply them. The results on the impact of the condition of anonymity were mixed.

Peer assessment and evaluation is a method of assessment of university students by their peers that has recently gained renewed pedagogical interest and broad implementation. It is an exercise in which students practice the skills needed for life-long learning (particularly, evaluation and critical thinking skills) by evaluating other students and observing how others evaluate the results of their learning.

Traditional assessment practices that preclude students' participation in the processes of evaluation thus perpetuating students' intellectual dependency are inconsistent with the revisited ideals and goals of the university education. Sharing with students the responsibility to participate in the assessment of their own and peers' work is a strategy that conforms to the conception of learning as active engagement and a dynamic ability to organize and modify ideas (Boud 1990; Zariski 1996). Any form of self and peer assessment has many potential benefits for the assessor and the assessee.[1] It encourages students' autonomy and higher order thinking skills. It augments students' responsibility for their own learning, intellectual independence, and self-confidence. It "helps students develop the ability to make judgments, a necessary skill for study and professional life" (Brown, Rust, and Gibbs, 1994).

The peer assessment data are frequently used in assigning individual students' grades. This raises a series of potential problems concerning the validity and reliability of peer evaluations, and questions about the merits and accuracy of the students' feedback. It also reminds us about the perils of intrusion into the private realm of students by making personal information publicly available.

The previous studies produced conflicting and inconclusive evidence of the reliability and validity of peer evaluations (Boud and Holmes 1995; Marcoulides and Simkin 1995; Penny and Grover 1996; Stefani 1995; see also Boud and Falchikov 1989, Oldfield and Macalpine 1995, and Stefani 1992). A review and meta-analysis of the studies of self-evaluations (Mabe and West 1982) demonstrated that self-evaluations were subject to a great deal of error resulting from the desire of self-enhancement and that people, in general, proved incapable of analyzing themselves objectively and reliably. Other concerns that were borne out in practice of peer assessment are gender, racial, and ethnic biases infiltrating the evaluation process (Layton and Ohland 2000).

The recurring tendency of students to bias their self- and peer-evaluations neutralizes the remarkable contributions of peer assessment to students' learning. To harness the peer assessment technique constructively, it is important to identify factors that jeopardize and/or enhance the validity and reliability of peer assessment. I conduct two experiments, in which I examine how the instrument of evaluation (with criteria vs. no criteria for evaluation provided), the format of evaluation procedure (anonymous vs. non-anonymous), and motivation of students (strong vs. weak) affect the accuracy of self and peer ratings.

Self and Peer Evaluations: When Students' Judgments are Flawed; Theory and Hypotheses

Self and peer evaluation is a type of judgment that students make about their own and their peers' academic performances. As any kind of social judgment, self and peer evaluations can be reasonably accurate or flawed because all human judgments differ in the amount of cognitive scrutiny they receive. When arriving at a conclusion or making a decision, people alternate between different modes of thinking. Sometimes they engage in careful, systematic, elaborate, processing of information to arrive at the best judgment possible (Kunda 1999, 235). On other occasions, people engage in more cursor, superficial, "quick and dirty," heuristic processing aimed at arriving at a satisfactory, if imperfect judgments (Kunda 1999, 235; Chaiken and Trope 1999). Thus, people alternate between the highly reasoned mode of thinking where available information is systematically reviewed, analyzed, and integrated prior to any judgment or decision and the intuitive superficial mode of thinking where judgments rely on relatively shallow situational cues, or on simple cognitive heuristics (Ajzen 1996, 300).

Although, both types of the modes of thinking can be invoked simultaneously, in a particular situation of making a judgment one mode will prevail depending on the person's *motivation* and *ability* to scrutinize evidence and process available information. As other individuals, students who have skills (ability) and motivation to examine critically other students' work will show less bias and more accuracy and consistency in their evaluations. Students do not usually acquire the same level of understanding of a subject matter compared to the teacher. This lack of familiarity with a domain of knowledge and the dearth of experience with judging other people's work may lead students to rely on different intellectual shortcuts and heuristics when making their judgments. The use of the criteria for evaluation will induce higher order thinking processes (application and analysis), thus, encouraging careful and guided reasoning. This leads to the following hypothesis:

Hypothesis I. The reliability of peer assessment improves when students are provided with instruments containing unambiguous criteria for evaluation.

When making responses, people are frequently guided by the considerations of social desirability, i.e., they tend to act in ways, which are perceived as acceptable to others. Publicity of judgments and responses may activate the social desirability heuristic: when acting in public, people do and say things, which they believe others approve of. When students make their evaluations publicly, social desirability can lead to inflated peer-evaluations because students may desire to be approved by other students. The fear of being deprecated, and the expectation of reciprocation from others may also lead to inflated assessments. Anonymity usually reduces the effects of social desirability leading to more honest answers and weighted solutions (Crowne and Marlowe, 1960; Joinson 1999). The hypothesis that follows is:

Hypothesis II. Anonymity of evaluating procedure improves the reliability of peer assessment.

When evaluating the performance of others, individuals often use their own performance as an anchor or a "yardstick" (Kunda 1999, 494) against which they measure the performance of others (Dunning and Cohern 1992; Dunning and Hayes 1996). Using self for judging others can substantially distort evaluations. The two primary types of self-bias extensively discussed in the literature are self-enhancement and downward comparison (Mabe and West 1982; Groeger and Grande 1996).

Self-enhancement is the unreasonably favorable self-appraisal that may be triggered by threats to self-esteem (Brown 1986; Markus and Kitayama 1991). When individuals are threatened by a superior performance by others, they will actively attempt to dispel the threat using an arsenal of strategies, such as downplaying the importance of the other's superior achievements, or underrating the performance of others (Kunda 1999, 499). The essence of the process of downward comparison is that persons can enhance their own subjective well being by comparing themselves with the less fortunate others (for a lucid discussion of downward comparison, see Wills 1981). The research on self and peer evaluations grants support to both self-enhancement bias and downward comparison (Layton and Ohland 2000). This suggests the following hypothesis:

Hypothesis III: Students' self-evaluations will be slanted toward higher appraisal of their own academic performance.

People's tendency to see themselves above average can be reduced when they are required to base their ratings on a small number of criteria supplied (Dunning et al. 1989). This suggests another hypothesis:

Hypothesis IV: Self-evaluations will be less biased when students are provided with the criteria for evaluations.

The social-psychological literature on self-evaluation suggests that self-enhancement and downward comparison may be associated with such factor as identification (vs. anonymity) of the rater. The researchers reported positive association of instructions of anonymity with accuracy of self-evaluations (Zariski 1996; Mabe and West 1982). I, too, anticipate finding that the condition ensuring anonymous marking will contribute to attenuation of self-enhancement bias.

Hypothesis V: Anonymity of evaluations increases the accuracy of self-reports.

Study I

Overview. To test hypotheses about the impact of evaluation instruments and anonymity of assessment on the reliability of peer evaluations I chose to conduct laboratory experiment. The integral features of experimental design, i.e., random assignment to conditions and organized manipulation of the independent variables, eliminate systematic error and provide better control over extraneous variables, thus making experiments superior to other research designs testing causal hypotheses.

The study was imbedded in the classroom curriculum. The students (participants of the experiment) were given a take-home assignment asking them to write a 1-1.5 page essay examining their understanding of the relevant concepts, principles, and theories learned in the class, and their ability to apply those theoretical constructs to the analysis of real-life situations. This assignment was later self and peer assessed.[2] Building the study in the course curricular and integrating it with the class routine made this experiment high on both mundane and experimental realism.

Subjects. I collected data from the undergraduate students enrolled in introductory political sciences courses at Purdue University during the Spring and Fall 2003 semesters. The sample of participants was very heterogeneous; students represented different academic backgrounds and levels of undergraduate education. Notwithstanding the size (N=70 in the first study, and N=40 in the second study) and the type (non-random, “convenient”) of the sample, I have reasons to believe that the reliability and validity of peer evaluations produced by the participants will not differ drastically from a larger sample of randomly drawn undergraduate students from the entire population of Purdue University undergraduates.

Design. The experiment followed a completely randomized between-subject 2 (instrument of evaluation: with criteria vs. no criteria provided) x 2 (anonymous vs. non-anonymous) factorial design. After the subjects submitted their essays, they were randomly assigned to one of four groups each receiving different treatment and/or level of treatment. On a day of the peer evaluation exercise, each student received a folder with the instructions page, a peer evaluation form, a personal essay, and the essays of four of the classmates (see Appendices I and II). The essays of the peers were distributed among participants at random. Each student had to evaluate 4 works (plus his or her own essay) and was evaluated by four other students and self-evaluated. The instructions page conveyed the rationale for participation in self and peer evaluation and succinct guidelines on how to do the evaluations.

Following the self and peer evaluation exercise, each student filled out an anonymous survey that asked them to rate how valuable they found the peer evaluation exercise for their personal learning of the material and whether they approved of incorporating peer assessment in the classroom curricular on a regular basis. The participants were also asked to identify the ways to improve peer assessment and the concerns that they had when completing the self and peer evaluation exercise. After conducting the experiment, the students were debriefed about the experimental purposes of the peer evaluation exercise and introduced to the results of the data analysis.

Experimental Manipulation. The evaluation forms and essays, the stimuli for the experiment, differed depending on the experimental conditions. To measure the impact of various evaluation instruments, I prepared two different peer evaluation forms, one containing no criteria or guidance on how the assessment should be carried out, and another one providing clear and informative criteria for the assessment of peers' works. The first type of the peer evaluation form was an adaptation of the peer evaluation instrument advocated by Brown (1995), in which students use a list of terms such as "excellent," "good," "marginal," and "unsatisfactory," to evaluate the completed assignments of their peers. The verbal ratings had a numerical equivalent for converting peer ratings into grades for the assignment. In the second peer evaluation instrument, each evaluative term was accompanied by a description of criteria to be used for assigning this evaluation and numeric rating to a peer's work (see Appendix II).

A half of the students received evaluation forms with explicit criteria for evaluation included; another half of the students were simply given an evaluation scale with no criteria for evaluation suggested. The expectation was that the evaluation instrument conveying the criteria for assessment would elicit higher order thinking processes (the application of the criteria to peers' work; the analysis of comprehensiveness of an essay and identification of the gaps in student's knowledge, etc.) and, therefore, lead to more reliable peer evaluation.

The second condition manipulated in the experiment was the anonymity of assessment. A half of the participants received essays marked with the names of their classmates, another half – with the unique identification numbers assigned to each student by the instructor (only the experimenter had access to the identification numbers of all of the participants). It was expected that the reliability of peer assessment would be higher under the condition of anonymous evaluation.

Dependent Variables.

(1) Reliability of Self and Peer Evaluations. In this study, I do not distinguish between the validity and reliability of peer evaluations. Precisely, validity of peer evaluations addresses the question of whether students evaluate what they suppose to evaluate, for instance, the intrinsic value of a student's contribution to a team task as opposed to an apparent "effort" or the amount of work undertaken by this student. The concerns with the validity of peer evaluations usually loom large in the collaborative learning. Working in the teams, student raters tend to evaluate the perceived academic ability rather than teamwork skills, and actual amount of work done rather than the built-in value of that labor (Stover 1976). The arrangements of peer evaluation exercise in this experiment discount this type of concerns with the validity of peer evaluations. That is why I concentrate on the reliability of peer assessment.

The reliability of peer evaluation refers to the extent to which peer assessment contains bias or variable errors, *i.e.*, errors that vary from one observation (assignment, paper, etc.) being assessed to the next and from time to time for a type of assignment evaluated twice or more by the same evaluation instrument. It is illustrative to view evaluation as composed of two components: a true component and an error component. The reliability of evaluation is, then, an index that summarizes discrepancies between the true scores and errors across a series of evaluations performed by a student using the same evaluation instrument. The problem with this measurement of the reliability of peer evaluation is that we don’t know what the “true” evaluation score is. In the studies of peer assessment, several substitutes for the “true” evaluation scores have been used, namely, instructor’s evaluations, current or mid-term (final) exam grades, and students’ GPA scores from the previous semester(s). Typically, the reliability of peer evaluation have been framed in terms of the ‘match’ (e.g., correlations, deviations, etc.) between marks students award themselves and their peers and the marks instructors would give for the same work (Zariski 1996, Mabe and West 1982). I used instructor’s evaluations as a substitute for the “true” scores.[3] I measure the reliability of peer evaluations as

$$\text{Reliability} = 3.5 - \left(\sum_{1}^{4} \mid \text{instructor's evaluation} - \text{peer evaluation} \mid \right).$$

Because discrepancies between students’ and instructor’s evaluations represented errors, I subtracted the sum of the discrepancies from a constant. This yielded an index of reliability in which higher scores reflected better (more reliable) peer evaluations.

(2) The Magnitude of Self-Bias. I was also interested in testing students’ evaluations on the presence of self-bias. The magnitude of the self-bias was measured by

looking at the difference in the deviations of the means of peer ratings from self-ratings, on the one hand, and instructor's ratings of the same students, on the other.

Results. The previous studies of peer evaluation demonstrated that students could under-evaluate, over-mark, or generate reliable assessments of the academic performances of their peers. This study demonstrates considerable variation in the reliability of peer evaluations. A total of 70 individual reliability indexes were examined. The variation in reliability ranges from 0 (a significant discrepancy [3.5 points on the scale of 5] between student's and an instructor's evaluations across four peer evaluations) to 3.5 (a 'perfect match' between student's and instructor's ratings) with the mean of 2.38 and standard deviation 0.85. Thus, the study reveals a considerable disparity between peers' and instructors' ratings in some cases as well as a great consistency of the peers' and instructors' grades in others.

I expected that students provided with criteria for evaluation would produce more reliable assessments of peers' essays. Independent-samples (Student's) t-test shows that peer evaluations based on the provided criteria are significantly more reliable than peer evaluations performed with the evaluation instrument containing no criteria for evaluation (M=2.59 with criteria, M= 2.17 with no criteria provided), $t_{70} = 2.11$, p=0.019. The evaluation instrument with explicit criteria for evaluation leads students to invest greater effort in making their judgments about peers' works.

It was expected that anonymous students' evaluations would be more reliable than non-anonymous ratings. Anonymous peer evaluations indeed turn out to be more reliable (anonymous M=2.5, non-anonymous M=2.25) but not quite significant, $t_{70} = 1.24$, p=0.10.

To assess whether there was an interaction effect of the instrument for evaluation and the condition of anonymity, I analyze the data using a 2x2 Analysis of Variance (ANOVA). While the test confirms the main effect of the instrument of evaluation ($F_{1,69}$=6.21, *p*=0.015) and anonymity of assessment ($F_{1,69}$=3.36, *p*=0.07), the interaction effect is not significant ($F_{1,69}$=2.22, *p*=0.14), i.e., the anonymous evaluations of students who also used criteria for ranking their peers' essays are not significantly improved relative to evaluations of students from other conditions.

Since peer evaluations are often accompanied by self-evaluation, the considerations regarding students' personal selves may color their judgments of others. To test for the biasedness of students' self-evaluations, I perform a one-sample (paired difference) t test of the null hypothesis that the mean of the magnitude of students' self-bias (y1) equals or is greater than that of instructor's (y2). In other words, I compare the deviations of the means of peer ratings from self-ratings, on the one hand, and instructor's ratings of the same students, on the other.

The students exhibite significant positive self-bias by rating their academic performance above that of their peers (the hypothesis that $y1 \geq y2$ can be rejected at 0.01 significance level ($p < 0.00005$)). A common concern with the inflation of self-evaluations is borne out in the study. The average self-rating is 4.64 and the average peer rating (the average of the means) is a significantly different 4.3 (on the scale from 0 to 5) compared to that of instructors' 4.12 and 4.3 correspondingly. Based on the sample at hand, 95% of the students tend to overate themselves by 0.2 – 0.5 points (0.5 is a whole latter grade on the scale from 0 to 5!). Only 18.5 percent of the students give themselves lower ratings than they give, on average, to others. And only 11 percent underrate themselves compared to the instructor's grades. Self-ratings of male and female students are not statistically different. The most disturbing feature of self-bias is that it does not disappear with changing the conditions of self and peer evaluation. The main effect for the condition of anonymity is not very significant ($F_{1,69}$=2.95, *p*=0.09]; the main effect for the evaluation instrument as well as effect of the interaction term are also statistically insignificant ($F_{1,69}$=0.09, p=0.77 and $F_{1,69}$=0.14, p=0.71 correspondingly). Thus, the analysis of variance does not reveal significant differences in the means of self-bias across the conditions of the experiment. Further studies about how to decrease self-bias are necessary.

The results of the study, generally, support advanced hypotheses. The evaluation instrument that offers explicit criteria for evaluation is proved to be very helpful in making peer evaluations more accurate. The anonymous evaluations are also more reliable, although this result is not highly statistically significant. The study demonstrates that students tend to exhibit significant positive self-bias by rating their academic performance above that of their peers. Self-bias is robust across experimental conditions, i.e., inflated self-evaluations do not disappear when students are provided with criteria for evaluation or when they do their evaluations anonymously.

Study 2

Overview. Study 2 preserves the essential conceptual features of Study 1 but strengthens manipulation with the instrument for evaluation to demonstrate how students can be further motivated to systematically process information and make better (more reliable) judgments of their peers' academic performance. The goal of Study 2 is to test the robustness of findings reported in Study I and provide further support to the theory suggesting that the level of motivation impacts the mode of thinking (superficial vs. in-depth) used by people in making their judgments. The criteria for evaluation enhance students' ability in making more reliable judgments about their peer performance. However, the paucity of strong motivation to apply the suggested criteria may not improve students' evaluation if students are motivated, instead, to arrive at a quick decision regarding the ratings of peers. The latter might occur because students often find contemplating over the application of the criteria so tedious that they are eager to get over it, or, because they are operating under time pressure and their thinking process "freeze" as soon as they arrive at what seems like a good enough solution (Kunda 1999, 242). There are different means to induce students to use the criteria in order to improve the accuracy of their evaluative judgments. For instance, the students may be motivated to systematically rely on the provided standards for evaluation by leading them to expect that the accuracy of peer evaluations will be evaluated (Kruglanski and Freund 1983). Study 2 tests the hypothesis about the impact of such a motivation induced by the announcement that students' assessments themselves will be evaluated. Additionally, Study 2 combines data from two experiments to test whether there is an increase in the reliability of peer evaluations carried out in the strong motivation condition compared to the conditions in which students were either simply given criteria for evaluation or given no criteria at all.

Study 2 was also imbedded in the classroom curriculum. The students were asked to prepare three 1-1.5 page essays targeting their comprehension of theoretical approaches to international relations and understanding of the relevant concepts, principles, and theories learned in the class. These essays were later self and peer evaluated. Participant of the experiment were introductory political science students at Purdue University (N=40). Peer evaluations were used for assigning individual grades for the assignment.

Design. The experiment followed a completely randomized between-subject 2 (strongly motivated vs. unmotivated) x 2 (anonymous vs. non-anonymous) factorial design. The procedure of Study 2 was identical to that of Study 1.

Experimental Manipulation. In the strong motivation condition, the students received evaluation forms with criteria for evaluation. To ensure that students attend to the information provided in the evaluation forms, the instructor read the criteria and their descriptions. To make students apply the criteria, the instructor promised bonus points to those students whose evaluations would be highly correlated with instructors' evaluations. The students in the weak motivation condition received neither criteria no promises of extra points for accurate evaluations.[4] The expectation was that in the strong motivation condition students' evaluation would be significantly more reliable.

There was no difference in manipulation of the condition of anonymity between Study 1 and Study 2. The dependent variable of interest, the reliability of peer evaluations, was defined and measured in the same way as it was done in Study 1.

Results. As expected, students strongly motivated to apply criteria when making judgments about their academic performance produced significantly more reliable peer evaluations (strongly motivated M=2.78, weakly motivated M=2.3), $t_{40} = 1.84$, $p = 0.036$. The results concur with theoretical predictions about the effects of motivation and abilities on cognitive processes. When strongly motivated and supported with guidance, students favor the elaborate over the cursory mode of thinking and processing of information, and are more likely to arrive at the most accurate and reliable judgment.

Students evaluating their peers under the condition of anonymity not only failed to outperform their peers from the non-anonymous condition but also produced less reliable evaluations. This is a statistically insignificant but contradictory to the findings of Study 1 result (anonymous M=2.66, non-anonymous M=2.41, $t_{40} = 0.94$, $p = 0.176$). An alternative to social desirability explanation of the effect of anonymity on individual judgments suggests that making individuals believe that their judgments will be made public can induce the highly reasoned mode of thinking (Kunda 1999, 238). Consequently, students who knew that their peers would be able to identify the raters might have been encouraged to make more careful and responsible judgments.

To assess the impact of peer evaluation instrument and strong motivation on the reliability of peer evaluations, I pooled the data from two studies, recoded the evaluation instrument variable (0 – no criteria provided, 1 – criteria provided, and 2 – criteria and strong motivation), and analyzed the data with the Ordinary Least Squares. To guard against the possibility that gender differences, variations in the experiences with peer evaluation, and levels of undergraduate education may account for the differences in reliability of peer ratings, controls for gender, experience with peer evaluation, and year at school were included in the model.[5] Table I presents the estimates of the model.

Table I. The Estimated Impact of the Evaluation Instrument and the Format of Evaluation on the Reliability of Peer Evaluations

Independent Variables	
Evaluation Instrument	0.216** (0.075)

Anonymity of Assessment	0.028 (0.16)
Gender	0.3* (0.16)
Experience	0.086 (0.06)
Year at School	-0.02 (0.07)
Constant	1.99 (0.26)

	N = 110
	R-squared = 0.12
	Adj. R-squared = 0.08

Note: Standard errors of regression are in parentheses.
**p < 0.1*; ***p < 0.01*

The coefficient on the evaluation instrument variable appears in the predicted direction and is highly statistically significant ($p < 0.01$). As hypothesized, evaluations of the students provided with no criteria for evaluation were, on average, less reliable, than the students' ratings based on the criteria. Furthermore, evaluations of the students strongly motivated to apply the criteria when making evaluative judgments were, on average, better relative evaluations of the weakly motivated students.

The test did not detect any significant differences in the reliability of evaluations completed under anonymous and non-anonymous conditions. Further investigation of the impact of the evaluation format (anonymous vs. non-anonymous) is necessary to determine whether social desirability effect or the fear of publicity influence students' judgments.

Among the control variables, gender was found to be slightly statistically significant. On average, males' evaluations tend to be more consistent with the instructor's evaluations than females' peer ratings. The coefficient on the experience variable appears in the predicted direction. More experience with peer evaluation is associated with more reliable ratings. The coefficient on the year at school variable is negative - a higher school level seems to be associated with less accurate evaluations. However, neither year at school nor experience was found to be statistically significant.

General Discussion and Conclusions

The new conception of learning focusing on the importance of life-long learning, metacognition, and student responsibility for their education (Zariski 1996) stirred up revisions in the teaching techniques and methods of assessment of students' academic performance. Peer evaluation tailors the revisited goals of university education by encouraging students' autonomy, intellectual independence, responsibility of their own learning, and higher order thinking skills. In a sample combining data from two studies, 67 percent of the students find the experience with evaluating their classmates' and their own work extremely or somewhat valuable for learning the material, practicing the skills of critical thinking, and self-appraisal (only 5% assign little or no value to peer evaluation). And 77% of the students from the sample favor the incorporation of practice of peer evaluation in the classroom curriculum.

Yet, the prospects of peer evaluation might be offset by the potential drawbacks with the lack of validity and reliability of peer assessment. This commonly expressed concern about self and peer ratings is borne out by the results of the study that reveal both a considerable unreliability of peer ratings in some cases as well as a notable consistency of peer evaluations in others. The students themselves recognize that they underrate or over-mark the works of their peers. In their comments on peer evaluation some students make the following remarks, "I feel I am too hard on others," "I thought some did not deserve the grade I gave them, it should have been lower," "I did not put effort into evaluation and gave random grades."

The promised benefits of peer assessment may only be realized after significant effort is made to incorporate it into our teaching practices in a way which is positive, non-threatening and attractive to students (Zariski 1996). The students in our sample express a number of concerns that might have had negative impact on the reliability of peer evaluations. For example, 47% of the participants of experiments express their worries about peers giving too low ratings; 34% are reluctant to give too low grade; 8% fear criticism; and 6% feel the lack of trust, respect, and rapport in the classroom. In their comments some students articulate their concerns with the validity of peer evaluations, "it's a controversial class, thus personal positions can lead to evaluation bias," "I felt it was an opinion paper. Not sure you can evaluate that. Some may not be able to distinguish," "if someone does not agree with the writer's stance, they may give an extremely low grade." Female students, on average, express more worries with peer evaluation than male students.

We can circumvent or minimize the potential problems of peer assessment by structuring the conditions of peer evaluations. The instrument of evaluation with criteria provided has positive effect on the accuracy of peer evaluations. This is a robust finding supported by both experiments conducted for the purposes of the study. Also, students strongly motivated to apply criteria produce more accurate peer evaluations compared to their peers provided with not criteria or supported with the criteria but not motivated to apply them.

The results on the impact of the condition of anonymity are mixed. In Study 1, anonymous evaluations were found to be more reliable, a result having moderate statistical significance (p<0.1); whereas in Study 2, students making their evaluations non-anonymously outperformed their peers from anonymous condition, a result, though statistically insignificant, contradictory to the findings from Study 1. Additional theorizing and tests are necessary to ferret out the confounding impact of social desirability and publicity on students' judgments.

The study reveals that students exhibite significant "self-enhancement" bias by rating their academic performance above that of their peers. The positive self-bias does not disappear with changing the conditions of self and peer evaluation.

The future studies of peer evaluations can test other means of inducing higher order thinking processes in students making evaluative judgments. For instance, the students may be motivated to systematically rely on the provided standards for evaluation by leading them to expect that after making their evaluations they would have to explain their thinking to others (Tetlock and Kim 1987). Researchers can look into the question of how valuable students' feedback is. Some teachers expressed doubts about the merits of formative evaluations. Here, the potential problems may concern both the content of such communication, their tone, and the ensuing effects on interpersonal relations and academic self-confidence (Zariski 1996). Again, a future study needs to address the question of whether the constructiveness of feedback received from peers depends on the application of appropriate criteria by students.

Another prospect area of research is self and peer assessment of the teamwork. Traditional education emphasized individualism, in contrast current academic practices increasingly involve team projects, cooperative learning, and an emphasis on the synergy possible though group processes (Van Duzer and McMartin 1999). Here, the researchers can explore the conditions for valid and reliable self and peer evaluation of the members of a team. Finally, a separate or imbedded into a larger project study can investigate gender and ethnic biases in peer evaluations.

Appendix I.

Peer Evaluation Form For Non-Anonymous Evaluations with No Criteria Provided[6]

Write down the names of the students whose works you received to evaluate (including your own) and next to each person's name write the word and grade from the following list that best describes that person's work:

Excellent (4.5 - 5 points)
Good (4.0 – 4.5 points)
Ordinary (3.5 – 4.0 points)
Marginal (3.0 – 3.5 points)
Unsatisfactory (0 – 3.0points)

Name	Rating

Date

Name of the rater

Appendix II.
Peer Evaluation Form with Criteria for Evaluation Provided

Write down the names of the students whose works you received to evaluate (including your own).[7]

Name					
Rating					

Next to each person's name write the word and grade from the following list that best describes that person's work. Make use of the criteria provided for evaluation:

The submitted assignment is:

Excellent (4.5 – 5 points)

- Shows a clear and complete understanding of the nature of diplomacy with recognition of the wider context within which it takes place.
- Applies key concepts, levels of analysis, and theories of international relations in the analysis of face-to-face leaders' diplomacy.
- Presents well organized, logical, cohesive, convincing, and coherent answer, which is easy to follow, flows well and have no internal inconsistencies.

Good (4.0 – 4.5 points)

- Demonstrates adequate understanding of the nature of diplomacy
- Shows understanding and application of key concepts, and levels of analysis.
- Contains an adequately organized and relatively easy to understand answer that avoids inconsistencies, and demonstrates good verbal skills.

Ordinary (3.5 – 4.0 points)

- Shows some understanding of diplomacy.
- Applies some key concepts of IR pertinent to diplomacy.
- Exhibits some skill in organizing and presenting information but with less clarity and elegance

Marginal (3.0 – 3.5 points)

- Demonstrates limited or incomplete understanding of diplomacy.
- Shows incomplete understanding of the relevant concepts.
- The answer is not easy to follow because of its poor organization and internal inconsistencies.

Unsatisfactory (less than 3.0 points)

- Demonstrates serious lack of understanding of diplomacy
- Exhibits no clear understanding of key concepts
- Lacks logic, coherence, and internal consistency, poorly organized and communicated ineffectively

Date **Name of the rater**

References

Ajzen, Icek. 1996. "The Social Psychology of Decision-Making" in *Social Psychology: Handbook of Basic Principles*, ed. T.E., Higgins, T.E and A.W. Kruglaski. New York: Guilford Press, 297-152

Biggs, J. 1999. *Teaching for Quality Learning at University*. Buckingham: SRHE and Open University Press

Boud, D. 1990. "Assessment and the promotion of academic values." *Studies in Higher Education* 15(1):101-11

Boud, David and Nancy Falchikov. 1989. "Quantitative studies of student self-assessment in higher education: a critical analysis of findings." *Higher Education* 18, 529

Boud, D. and Holmes, H. 1995. "Self and peer marking in a large technical subject." in *Enhancing Learning through Self Assessment*, ed. D. Baud. London: Kogan Page, 63-78

Brown, Jonathon D. 1986. "Evaluations of self and others: self-enhancement biases in social judgments." *Social Cognition* 4(4):353-376

Brown, S., Rust, C., and Gibbs, G. 1994. "Involving Students in the assessment process." in *Strategies for Diversifying Assessment in Higher Education*, ed. S. Brown, Rust C., and Gibbs, G. Oxford: Oxford Center for Staff Development, 21-24

Brown, R.W. 1995. "Autorating: Getting individual marks from team marks and enhancing teamwork." In *Frontiers in Education Conference Proceedings*. Pittsburgh, IEEE/ASEE. November. <http://fie.engrng.pitt.edu/fie95/3c2/3c24/3c24.htm>. (September 20, 2003)

Chaiken, S., and Trope, Y. 1999. *Dual-Process Theories in Social Psychology*. New York: Guilford

Crowne, D. P. and D. Marlowe, D. 1960. "A new scale of social desirability independent of psychopathology." *Journal of Consulting Psychology* 24, 349-354

Dunning, D. and Cohen, G.L. (1992). "Egocentric definitions of traits and abilities in social judgment." *Journal of Personality and Social Psychology* 63, 341-355

Dunning, D. and Hayes, A.F. (1996). "Evidence for egocentric comparison in social judgments." *Journal of Personality and Social Psychology* 71, 213-229

Groeger, J.A. and Grande, G.E. 1996. "Self-preserving assessment of skill?" *British Journal of Psychology* 31(4): 61-79

Heywood, J. 2000. *Assessment in Higher Education*, London: Jessica Kingsley Publishers

Joinson, A.N. 1999. "Social desirability, anonymity and Internet-based questionnaires." *Behavior Research Methods, Instruments and Computers* 31(3):433-438

Kaufman, Deborah B., Richard M. Felder, and Hugh Fuller. 1999. "Peer Ratings in Cooperative Learning Teams." In *ASEE Annual Conference Proceedings*, ASEE, Charlotte. June.

Kruglanski, A.W and Freund, T. 1983. "The freezing and unfreezing of lay-inferences: Effects on impressional primacy, ethnic stereotyping, and numerical anchoring." *Journal of experimental Social Psychology* 19, 448 – 468

Kruglanski, A.W. and D.M. Webster. 1996. "Motivated closing of the mind: "Seizing" and "freezing." *Psychological Review* 103, 263-283

Kunda, Ziva. 1999. *Social Cognition: Making Sense of People*. A Bradford Book, The MIT Press, Cambridge, MA

Layton, Richard A. and Matthew W. Ohland. 2000. "Peer Evaluations in Teams of Predominantly Minority Students." In *American Society of Engineering Education Proceedings*, Session 2330, ASEE, Washington, DC

Mabe, P.A. and West, S.G. 1982. "Validity of self-evaluation of ability: A review and meta-analysis." *Journal of Applied Psychology*, 67(3):280-296

Marcoulides, G.A, and Simkin, M.G. 1995. "The consistency of peer review in student writing projects." *Journal of Education for Business* 70(4):220-223

Marcus, H. and Kitayama, S. 1991. "Culture and the self: Implications for cognition, emotion, and motivation." *Psychological Review* 98, 224-252

Occhipinti, John D. 2003. "Active and accountable: teaching comparative politics using cooperative team learning." *Political Science and Politics* 36 (January):69-75

Ohland, Matthew W. and Richard A. Layton. 2000. "Comparing the Reliability of Two Peer Evaluation Instruments." <http://www.succeed.ufl.edu/papers/00/00072.pdf> (October 3, 2003)

Oldfield, Keith A. and Macalpine, Mark K. 1995. "Peer and self-assessment at tertiary level – an experiential report." *Assessment and Evaluation in Higher Education* 20(1):125-32

Penny, A.J. and Grover, C. 1996. "An Analysis of Student Grade Expectations and Marker Consistency." *Assessment & Evaluation in Higher Education* 21(2):173-183

Stefani, L.J. 1992. "Comparison of collaborative self, peer and tutor assessment in a biochemistry practical." *Biochemical Education* 20(3):148-51

Stefani, L.J. 1995. "Peer, self and total assessment: relative reliabilities." *Studies in Higher Education* 19(1):69-75

Stover, Robert V. 1976. "The impact of self-grading on performance and evaluation in a constitutional law course." *Teaching Political Science* 3(3):303

Tetlock, P.E. and Kim, J.I. 1987. "Accountability and judgment processes in a personality prediction task." *Journal of Personality and Social Psychology* 52, 700-709

Wills, Thomas Ashby. 1981. "Downward Comparison Principles in Social Psychology." *Psychological Bulletin* 90(2):245-271

Zariski, Archie. 1996. "Student Peer Assessment in Tertiary Education: Promise, Perils and Practice." In *Teaching and Learning Within and Across Disciplines*, ed. J. Abbot and Willcoxson, L. Proceedings of the 5th Annual Teaching Learning Forum, Murdoch University. February. Perth: Murdoch University. 189-200 <http://cea.curtin.edu.au/tlf/tlf1996/zariski.html> (October 10, 2003)

[1] There is a variety of forms of self and peer evaluation including, but not limited to, formative peers' reviews to provide feedback, summative grading, evaluation as an element of students' tutoring, etc. Peer assessment may include the prior setting of criteria and the selection of evidence of achievement and can be used in conjunction with collaborative learning or by itself (Biggs 1999; Brown, Rust and Gibbs 1994; Occhipinti 2003). Peer-assessment is often combined or considered together with self-assessment.

[2] Peer assessment can be structured in a wide variety of ways and the literature records many permutations (Zariski 1996). I chose to offer our students to evaluate their written assignments because team projects were not the part of the classroom curriculum and

collaborative learning was not among the objectives of our courses. "Logistical" problems, e.g., big classes and stationary seats in the classrooms, also shrank the range of our choices of the exercises that could be used for self and peer evaluation.

[3] In grading students' assignments I relied on the same guidelines offered to students in the condition of evaluation instrument with criteria for evaluation provided. To ensure unbiasedness of instructor's evaluations, I asked a colleague teaching a different section of the same course to grade a random subset of students' essays. The correlation coefficient of two instructors' rankings was 0.8.

[4] During the assignment of actual grades for essays, students in all conditions received bonus points if their evaluations were highly correlated with instructor's evaluations.

[5] Gender is a dichotomous variable with *male* coded as 1 and *female* – as 0. The experience with peer evaluation is coded on the scale from 1 (a great deal of experience) to 6 (none). *Year at school* ranged from 1 to 4 (1 – freshmen; 2 – sophomore; 3-junior; and 4 - senior).

[6] The participants in the anonymous evaluation condition were asked to write down the identification numbers of the students whose works they received to evaluate (including their own).

[7] See footnote 6.