The Journal of Economic Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/vece20

Gender and peer evaluations

Molly Espey

To cite this article: Molly Espey (2021): Gender and peer evaluations, The Journal of Economic Education, DOI: 10.1080/00220485.2021.2004277

To link to this article: https://doi.org/10.1080/00220485.2021.2004277

https://doi.org/10.1080/00220485.2021.2004277

Routledge
Taylor & Francis Group

RESEARCH IN ECONOMIC EDUCATION

Gender and peer evaluations

Molly Espey

Department of Economics, Clemson University, Clemson, SC, USA

ABSTRACT

Peer evaluations for nearly 1,650 students in three different team-based learning economics courses are analyzed for evidence of gender differences in ratings given and received. The analysis controls for general academic skills and economics-specific skills as well as other individual and team characteristics. Females earn higher evaluations than males in introductory and intermediate-level microeconomic theory courses even when conditioning on achievement, but there is no statistically significant difference in a more advanced course. Finally, there is little evidence of gender bias in evaluations given; both males and females rate female teammates higher than male teammates.

KEYWORDS

Economic education; gender; peer evaluation; team-based learning

JEL CODES

A2; A22

"Whether women are better than men I cannot say—but I can say they are certainly no worse."

—Golda Meir

Are women better teammates than men? I analyze cooperation among students in college-level classroom teams to assess gender differences in evaluations received by teammates. Teamwork is an increasingly common aspect of the modern economy (Deloitte 2016). Learning to work more productively in teams can benefit college students in a variety of ways, improving their grades, increasing their learning, and enhancing skills that can be valuable in their future careers.

In this study, teams are assigned by the instructor, and students work in the same team throughout the semester each class period on both graded and ungraded in-class activities. Students have the opportunity to evaluate their teammates twice during the semester: midterm and at the end of the semester. Evaluations are intended to measure students' contributions to team interactions, including attendance, preparation, and respect for others. They are designed in a manner to minimize free-riding by giving each student a set number of points to allocate to his/her teammates and requiring that at least one score be differentiated.

A variety of incentives exist to reward positive teamwork. Effort in team activities contributes to an individual student's grade through an overall team grade, peer evaluations that are used to weight the team grade for each individual, and individual learning. Students report improvements in learning from a team-based environment relative to a lecture-based environment (Espey 2018b), and those who earn higher peer evaluations from their teammates earn higher grades on individual exams, even when controlling for entering academic ability (Espey 2018a).

Women are estimated to earn about 4.2 percent higher peer evaluations on average than men in an introductory-level course and about 3.4 percent higher in an intermediate-level course when controlling for other explanatory factors. While this may not seem like much, it can translate to 2 to 4 percentage points more for the peer evaluation-weighted team portion of an individual's grade. Combined with the additional learning associated with greater effort on team activities, this is easily enough to raise an individual's grade from a high C to a low B, or from a high B to a low A.

CONTACT Molly Espey mespey@clemson.edu Professor, John E. Walker Department of Economics, 312H Wilbur O. and Ann Powers Hall, Clemson University, Clemson, SC 29634, USA.

Interestingly, the gender difference in evaluations is smaller in the intermediate-level course than in the introductory-level course and disappears entirely in a more advanced course that has the intermediate-level course as a prerequisite. This supports developmental literature that males are slower to mature and to develop the interpersonal skills to work productively in a cooperative environment, but it also could reflect differences in opportunity cost by gender in the earlier years of college. Additionally, this result could be due to self-selection into a higher-level economics course that reduces other differences between males and females; this study does not attempt to ascertain the reason for these differences.

In previous research, both age and context have been found to play a role in gender differences in individual contributions to team or cooperative activities. Jacob (2002) found eighth-grade girls had higher noncognitive skills than boys did. Among these noncognitive skills are organizational skills, the ability to pay attention, and communication skills, all characteristics related to being a more valuable or productive team member. Babcock and Laschever (2003) found many women display cooperative personality traits that give them a comparative advantage in contexts where cooperation provides mutual gains. However, the experimental evidence on this, based on public-good games, ultimatum games, and dictator games, is mixed (Niederle 2016).

Additionally, personality traits are not unchanging. The greatest mean-level change in personality traits after childhood takes place in young adulthood (Almlund et al. 2011). This includes traits such as agreeableness, conscientiousness, and emotional stability that would be expected to correlate to more positive and productive interactions with others.

Rather than focusing on individual factors influencing peer evaluations, previous research has tended to focus on the credibility or accuracy of student ratings in comparison to instructor ratings (for example, Falchikov and Goldfinch 2000; Wever et al. 2011) or on learning or behavioral outcomes of the process. For example, Liu, Pysarchik, and Taylor (2002) reported peer evaluations encourage individual responsibility, while Topping et al. (2000) suggested the process enhances students' meta-cognitive perceptions. Lee and Lim (2012) find students indicate social contributions, such as organizing or coordinating managerial abilities, were more important than cognitive contributions in their evaluation of peers. The research presented here is a unique quantitative analysis of how student peer evaluations during a semester-long interaction in a team-based learning environment are related to individual student characteristics.

This research complements existing literature on evaluations and gender that includes analysis of performance reviews, tenure and promotion, and student evaluations of instructors. Igbaria and Baroudi (1995) found no difference between men and women in performance reviews in the information systems workplace. In contrast, Lyness and Heilman (2006) found managers in financial services rated women lower than men, and those women who were promoted had higher performance evaluations than males who were promoted, suggesting women were held to a higher standard. Similarly, in his study of large service organizations, Castilla (2008) found women and minorities were rewarded less for equivalent evaluations received by nonminority males. In a study of U.S. universities, Weisshaar (2017) determined the gender gap in promotion and tenure was not fully explained by discipline or productivity. De Paola and Scoppa (2015) found females were less likely to be promoted by an exclusively male committee in their study of Italian universities.

Numerous analyses of student evaluations of teaching have been conducted, with conflicting results regarding gender. For example, both Bachen, McLoughlin, and Garcia (1999) and Basow (1995) determine that female students rate female instructors higher than male instructors, while male students rated male and female instructors similarly. In a comparative study of French and U.S. institutions, Boring, Ottoboni, and Stark (2016) found male students in France rated male instructors higher than female instructors, while in the United States, the gender difference in evaluation was primarily due to female students rating male instructors higher than female instructors. Finally, both Arbuckle and Williams (2003) and MacNell, Driscoll, and Hunt (2015) conducted experiments using online courses with hidden or disguised instructor gender identity and determined bias in favor of instructors represented as male regardless of their true gender.

Contrary to most of these previous studies, I do not find evidence of gender bias against females in this analysis of peer evaluations. Both men and women rate women significantly higher than they rate men in the introductory and intermediate-level courses, with no statistically significant difference

between their ratings in a more advanced course. Further, women do not appear to be disadvantaged by being the only woman or only one of two women on a team, with no statistically significant impact on peer evaluations in such situations.

In this study, I first provide an overview of team-based learning (TBL) and then present the analysis of thirteen years' worth of TBL in introductory microeconomic theory classes, six years in intermediate microeconomics, and nine years in environmental economics to determine how individual and team characteristics contribute to peer evaluations. The study adds to the literature by focusing on peer evaluations in a cooperative environment with repeated interactions. I also analyze gender differences in peer evaluations in terms of overall differences between how males and females are rated, how these differences vary depending on the setting, and finally, whether or not there is a difference between how males and females rate each other.

Experimental setting: Team-based learning

Team-based learning (TBL) is a student-centered instructional strategy that consists of modules taught in a three-step cycle: preparation, in-class readiness assurance testing, and application-focused exercises. Teams of five to seven students are created by the instructor to ensure diversity and a reasonably even distribution of academic assets across teams based on incoming grade point average (GPA) and major. Students are individually responsible for readiness assurance tests (RATs) based on preparatory reading at the beginning of each of five modules or units over the semester, individually completed homework assignments, and exams. Individuals are responsible to teams through team RATs comprising the same questions as the individual RATs and through daily team application activities.

Each unit of the course begins with an individual RAT, followed by a team RAT over the same questions. The team RAT uses a scratch-off sheet that allows teams multiple attempts such that by the end of the test, they know the correct answer to all of the questions. Team activities vary from basic multiple choice questions on the board over beginning applications, to brainstorming, to more complex analysis of real-world applications toward the end of each module. Most activities are designed to follow the "4S" structure: significant problems, same problem for all teams, specific choice, and simultaneous reporting (Michaelsen, Knight, and Fink 2002).[1] Significant problems engage students in concrete examples so that they understand the usefulness of the course concepts. Specific choices require teams to take a position, sometimes also requiring them to support that position with a short rationale for their choice. Forcing all students to confront the same problem enables them to better engage with each other across teams, while simultaneous reporting precludes teams from simply agreeing with the majority of others, forcing them to decide before knowing what other groups will say. Students are given time to address the problem in their teams and then report their answers. Whole-class discussion is used to address differences across teams and clarify any uncertainties. Each team has a folder in which the instructor distributes team and individual assignments, and teams turn in completed work at the end of each class period.

Peer evaluations hold individuals responsible to their team and minimize free riding. Each student has total points equal to 10 times the number of teammates he/she has, must allocate all of the points, can assign only whole number evaluations, and must differentiate at least one score. For example, on a team of six, each student has 50 points to allocate to five teammates. A student could give one 12, two 10 s, and two 9 s, or one 11, three 10 s, and one 9, or any other combination that adds up to 50 except giving everyone a ten. Specifically, students are provided with the following directions:

> "Please assign scores that reflect how you really feel about the extent to which the other members of your team contributed to your learning and/or your team's performance. Consider such things as:
>
> - Attendance – Do they attend class consistently?
> - Preparation – Were they prepared when they came to class?
> - Contribution – Did they contribute productively to group discussion and work?
> - Respect for others – Did they encourage others to contribute and listen to the ideas of others?

It is important that you raise the evaluation of those who truly worked hard for the good of the group and lower the evaluation of those you perceived did not work as hard on group tasks."

Evaluations are completed mid-semester, but only the end-of-semester peer evaluation counts toward the final grade; it is used to weight the team grade, which is 25 percent of a student's overall grade. If a student averages 10 points on the evaluations from his peers, for example, and his team is averaging 88 percent, the team portion of that student's grade is 88 percent. In contrast, if he averages 10.5 on his peer evaluations, the team portion of his grade will be 92.4 percent (105 times 0.88).

Data

Evaluations of more than 1,600 students on 282 teams in three different courses, taught between 2007 and 2019 at a four-year public U.S. university, were analyzed to determine what quantifiable characteristics of individuals and teams influence students' ratings of each other. The analysis includes 22 sections of introductory microeconomics offered between 2007 and 2019, 13 sections of intermediate microeconomics offered between 2014 and 2019, and 10 sections of environmental economics, which has a prerequisite of intermediate microeconomics, offered between 2011 and 2019. Each student is evaluated by five teammates on average, resulting in more than 8,000 evaluations.

Summary statistics by course by gender are shown in table 1. Females make up close to 42 percent of the introductory-level course, 36 percent of the intermediate-level course, and almost 30 percent of the environmental economics course. Females are younger on average than males in both the introductory and intermediate-level courses, with proportionately more freshman females (including first-semester freshmen) and more sophomore males in the introductory course, and proportionately more sophomore females and more senior males in the intermediate-level course.

Females have a statistically significantly higher average GPA when entering all three classes. Yet, there is not a significant difference in performance between males and females on the first readiness assessment test (RAT 1). The final exam score is not used in the analysis, although it is included for comparison purposes and shows that final performance in the course is not significantly different between males and females in the intermediate-level course, but females score slightly higher than males in the other two courses (significant at the 5% level).

Peer evaluations of females average slightly over 10.2 for all three courses, ranging from 0 to 20 in both the introductory and intermediate-level courses and from 5 to 15 in the environmental course. Evaluations for males average slightly over 9.7 in the introductory-level course, ranging from 0 to 16, and average just under 9.9 in the other two classes, ranging from 0 to 18 in the intermediate-level course and 0 to 17 in the environmental course. Evaluations are typically in the 7 to 12 range, however, with less than five percent falling outside this range.[2]

Team characteristics are shown in table 2. In the regression model, rather than including team average GPA, the team average omitting the individual observed is included. Hence, if the individual observed has a higher GPA than his/her team, the "left out average GPA" will be lower than the team average, and if the individual has a lower GPA than the team's average, the "left out average GPA" will be higher. The percent female on a team ranges from 0 to 100 for the intro class, 0 to 80 in the intermediate class, and 0 to 75 percent for the environmental class. Other measures of gender diversity within teams were assessed but did not affect the results.[3]

The class diversity variable is constructed similarly to an inverse Herfindahl-Hirschman index or Simpson's diversity index, equal to 1 minus the sum of squared proportions of each class level (freshman, sophomore, junior, and senior), such that a team comprising only one class level has a diversity value of 0 and a team with equal proportions of each class level has a diversity value of 0.75. Class diversity ranges from 0 to the max of 0.75 for the introductory class, 0 to 0.72 in the intermediate class, and 0 to 0.67 (which is the max for that class because there are no freshmen in it).

As TBL guidelines recommend that teams range from five to seven in size, all but eleven teams in this study fall within that range, with an overall average size of about six. In the exceptions to this, students withdrew too far into the semester to make team size adjustments, resulting in teams of four.

Table 1. Summary Statistics for Individual Characteristics

	Introductory Micro		Intermediate Micro		Environmental	
	Female (n = 366)	Male (n = 512)	Female (n = 160)	Male (n = 283)	Female (n = 96)	Male (n = 229)
Peer evaluation	10.27 (1.44)	9.79 (1.64)	10.29 (1.34)	9.84 (1.54)	10.25 (1.32)	9.90 (1.66)
RAT1 (out of 60)	47.89 (7.77)	47.96 (7.91)	44.53 (7.86)	44.35 (8.57)	43.53 (7.75)	42.85 (8.53)
GPA[a]	3.099 (0.576)	2.808 (0.667)	3.283 (0.580)	3.024 (0.616)	3.297 (0.564)	2.965 (0.615)
Final exam	0.814 (0.099)	0.797 (0.110)	0.797 (0.132)	0.788 (0.128)	0.808 (0.096)	0.781 (0.112)
First-semester freshman	15.6%	10.5%	–	–	–	–
Transfer or international[b]	5.5%	7.8%	1.9%	7.8%	4.2%	1.7%
Freshman	29.9%	24.6%	2.5%	3.2%	–	–
Sophomore	37.5%	41.0%	35.6%	32.2%	5.2%	6.1%
Junior	22.5%	21.9%	43.8%	44.2%	36.5%	36.7%
Senior	10.1%	11.9%	18.1%	19.8%	58.3%	57.2%
Out-of-state	35.8%		45.0%		47.9%	
Lone female	5.2%	24.8%	10.6%	32.9%	14.6%	46.7%

Note: Standard deviations are in parentheses.
[a]GPA is not available for first-semester freshmen, new transfer students, or international exchange students.
[b]All of the students in this category in the environmental economics course are international exchange students.

Table 2. Summary Statistics for Team Characteristics

	Introductory Micro	Intermediate Micro	Environmental
	(n = 151)	(n = 74)	(n = 57)
Team GPA	2.929 (0.197)	3.144 (0.214)	
Percent female	41.8 (19.5)	36.2 (16.0)	29.7 (18.8)
Class diversity	0.568 (0.128)	0.507 (0.142)	0.441 (0.133)
Team size	5.894 (0.767)	5.892 (0.885)	5.684 (0.805)

Note: Standard deviations in parentheses.

Introductory-level classes ranged in size from 19 to 60 students, with an average of 43, intermediate-level classes ranged from 26 to 60 (except for three summer sections of fewer than ten students) with an average of about 48, and environmental classes ranged from 22 to 38, averaging about 32.

Rater characteristics are reflective of individual characteristics. However, the number of times each gender has the opportunity to rate a teammate of the same or different gender depends on the gender composition of the team. Because one focus of this analysis is to determine if there are gender biases in evaluations, interaction variables are constructed for the gender of the individual being rated crossed with the gender of the evaluator. These are indicated in table 3, which shows the proportion of overall evaluations by gender mix. Thus, "F rated by F" indicates a female evaluated by a female teammate, "F rated by M" indicates a female evaluated by a male teammate, "M rated by F" indicates a male evaluated by a female teammate, and "M rated by M" indicates a male evaluated by a male teammate. In all three courses, males have more opportunities to evaluate males, and females have fewer opportunities to evaluate females simply because the gender composition of the courses was predominantly male in all cases. As male evaluations of females and female evaluations of males are reciprocal, these proportions are equal for each course. Note that there is some discrepancy between the number of students who completed a course and the number of students who completed evaluations. These students are included in the analysis in that they were evaluated by their teammates, but they are missing as evaluators because they did not submit peer evaluations.

Model

The peer evaluation received by an individual should reflect that student's general academic abilities, economics-specific academic abilities, his/her outside-of-class preparation for in-class engagement, as well as noncognitive skills related to working with others such as interpersonal communication skills, attitude, focus, and attention. General academic abilities are assumed to be reflected in overall

Table 3. Individual by rater gender breakdown.

	Introductory Micro (n = 4383)	Intermediate Micro (n = 2237)	Environmental (n = 1565)
F rated by F	16.7%	11.1%	8.1%
F rated by M	24.6%	24.5%	21.4%
M rated by F	24.7%	24.5%	21.4%
M rated by M	34.0%	39.9%	49.1%

GPA, while economics-specific skills should be reflected in the first RAT score. The first RAT covers basic introductory concepts such as supply and demand in the introductory-level course, a review of basics including elasticities in the intermediate-level course, and some core environmental economics concepts such as externalities, efficiency, and market failure for the environmental economics course. Further, the first RAT is given on the second or third day of class before any significant interaction with future teammates. Both GPA and the RAT1 score will also reflect individual effort outside of the classroom.

Each individual is evaluated by each of his/her teammates, so there are multiple observations for each individual relating to the evaluations given by their teammates. Each individual evaluation is modeled as a function of characteristics of that individual (I_i), characteristics of that person's team (T_i), and class controls (C). Gender is of particular interest in terms of the raters, so variables indicating gender of the individual being evaluated interacted with the rater's gender are also included in the vector R. Other rater characteristics are not included as each student has a set number of points to allocate to his/her teammates so cannot systematically rate differently except in interaction with individual characteristics. The evaluation of student i by each teammate j is:

$$Eval_{ij} = \alpha + I_i\beta + T_i\tau + C_i\lambda + R_{ij}\delta + \varepsilon_{ij}$$

where the vector I includes individual i's gender, GPA, RAT1 score, and indicator variables for class level, first-semester freshman, transfer student, and out-of-state student.[4] The vector T includes left-out-average GPA (GPA of the other individuals on a given student's team), percent female on the team, class diversity index, and team size. The vector R includes the interaction variables between the gender of the individual evaluated and the gender of the evaluator (F rated by F, F rated by M, M rated by F) with males rated by males as the base.

Empirical results

Empirical results are shown in tables 4, 5, and 6 for the introductory, intermediate, and environmental economics courses, respectively. In all three courses, both GPA and RAT1 scores are positive and significant, implying that both general and course-specific academic skills contribute positively to evaluations by others. One's academic abilities in relation to others on one's team ("leftoutgpa") is also statistically significant and negative, indicating those who have higher grade point averages than their teammates (lower "left out gpa") tend to be rated more highly than those with lower GPAs than their teammates.

Controlling for other factors, females earn almost half a point higher peer evaluations in the introductory-level course (0.42) and about a third of a point higher in the intermediate-level course (0.34). There is no statistically significant difference between male and female evaluations earned in the environmental economics course. This lends credence to the idea that females have more noncognitive skills, such as communication and interpersonal skills, when entering college, which would contribute to more successful engagement within a team environment in the classroom. However, the smaller size of the coefficient in the intermediate-level course and the lack of difference in the predominately senior-enrollment environmental course suggests that perhaps males catch up over time in college. This also could be related to self-selection, with the vast majority of the students enrolled in the environmental economics course being either economics majors or minors. As a different, yet related, issue, there is often concern about an individual being the only female on a team, whether in the classroom or in a work environment. However, including an indicator variable for a woman being the only female (or one of two) on a team was not statistically significant for any of the courses.

Table 4. Empirical Results for Introductory Microeconomics Peer Evaluations (N = 4347)

	Model 1	Model 2	Model 3	Model 4
Individual Characteristics				
Female	0.371** (0.048)	0.421** (0.052)	0.420** (0.052)	
F rated by F				0.606** (0.083)
F rated by M				0.391** (0.064)
M rated by F				0.075 (0.063)
GPA	0.469** (0.037)	0.406** (0.038)	0.394** (0.042)	0.392** (0.042)
RAT1 score	0.1028** (0.003)	0.028** (0.003)	0.031** (0.003)	0.031** (0.003)
Team Characteristics				
Percent female		−0.407** (0.131)	−0.400* (0.152)	−0.568** (0.164)
Left out GPA		−0.535** (0.099)	−0.557** (0.136)	−0.557** (0.136)
Class diversity		0.232 (0.207)	0.260 (0.241)	0.257 (0.240)
Team size		−0.018 (0.034)	0.010 (0.044)	0.011 (0.044)
Other indiv controls	Y	Y	Y	Y
Class controls			Y	Y
Adjusted R-squared	0.0934	0.1012	0.0999	0.1013

Note: Standard errors are in parentheses.
*$p < .05$; **$p < .01$

Table 5. Empirical Results for Intermediate Microeconomics Peer Evaluations (N = 2201)

	Model 1	Model 2	Model 3	Model 4
Individual Characteristics				
Female	0.328** (0.065)	0.342** (0.068)	0.341** (0.068)	
F rated by F				0.542** (0.114)
F rated by M				0.275** (0.083)
M rated by F				0.014 (0.080)
GPA	0.451** (0.053)	0.437** (0.054)	0.415** (0.062)	0.413** (0.063)
RAT1 score	0.028** (0.004)	0.027** (0.004)	0.028** (0.004)	0.027** (0.004)
Team Characteristics				
Percent female		−0.345 (0.211)	−0.386 (0.243)	−0.517* (0.256)
Left out GPA		−0.687** (0.115)	−0.805** (0.201)	−0.828** (0.201)
Class diversity		0.246 (0.242)	0.224 (0.255)	0.231 (0.255)
Team size		−0.024 (0.041)	−0.046 (0.057)	−0.051 (0.057)
Other indiv controls	Y	Y	Y	Y
Class controls			Y	Y
Adjusted R-squared	0.1090	0.1255	0.1211	0.1226

Note: Standard errors are in parentheses.
*$p < .05$; **$p < .01$

The reason that the coefficient on the percent of the team being female is negative is likely due to the facts that most of the teams are majority male and how points can be allocated is restricted. If females are truly better teammates and deserving of higher peer evaluations, given that each person has only a limited number of points to allocate to his/her teammates, the more females there are on a team, the more people there are who may be assigned above average evaluations. For every person receiving an evaluation above 10, someone must be below 10. Notably, for the one course in which there is not a statistically significant difference between evaluations for men and women, the percent female on the team also is not statistically significant.

The interaction variables for the gender of the evaluated person crossed with the gender of the rater are included in Model 4 to address the possibility of gender bias in evaluations. The omitted variable is the most common situation of males rated by males. For both the introductory and intermediate-level courses, females are rated higher than males on average by both female and male teammates. Female ratings of females average about 50 percent higher than males' ratings of females in the introductory-level course and are nearly twice as high in the intermediate-level course; both differences are statistically significant at the 5 percent level. Yet, in the environmental economics course, none of the gender interaction terms is statistically different from zero, reinforcing the results of the other models' finding of no significant difference between male and female evaluations in that course.

Table 6. Empirical Results for Environmental Economics Peer Evaluations (N=1526)

	Model 1	Model 2	Model 3	Model 4
Individual Characteristics				
Female	0.024 (0.082)	0.028 (0.089)	0.027 (0.089)	
F rated by F				0.140 (0.154)
F rated by M				0.035 (0.106)
M rated by F				0.074 (0.102)
GPA	0.954* (0.070)	0.845* (0.074)	0.848* (0.084)	0.849* (0.084)
RAT1 score	0.017* (0.005)	0.019* (0.005)	0.020* (0.005)	0.020* (0.005)
Team Characteristics				
Percent female		−0.063 (0.230)	−0.109 (0.254)	−0.207 (0.276)
Left out GPA		−0.925* (0.189)	−0.899* (0.265)	−0.896* (0.265)
Class diversity		0.018 (0.343)	0.020 (0.358)	0.018 (0.358)
Team size		0.002 (0.057)	0.002 (0.061)	0.003 (0.061)
Other indiv controls	Y	Y	Y	Y
Class controls			Y	Y
Adjusted R-squared	0.1856	0.1998	0.1959	0.1953

Note: Standard errors are in parentheses.
*$p < .01$

Conclusions

Women earn higher peer evaluations than males in a team-based learning environment in introductory and intermediate microeconomic theory courses, but there is not a statistically significant difference in an upper-division environmental economics course. Further, both males and females rate females higher in these first two courses, with no statistically significant difference between the genders in terms of how they evaluate men.

Students earn higher evaluations for a variety of reasons, including for being more academically prepared, working harder outside the classroom in preparation for interaction with peers, being more engaged with others during class (focusing on the task, not on their phone or computer, communicating with all team members), volunteering to write, encouraging others to provide input, and listening to others. Previous research (Espey 2018a) indicates that regardless of their initial academic abilities, students who earn higher peer evaluations learn more than those who earn lower peer evaluations. However, previous research also indicates that all students, regardless of GPA or gender, feel that team-based learning improves their critical thinking and problem-solving skills better than a lecture environment does (Espey 2018b). It is also notable that male and female peer evaluations converge in the more advanced course, reinforcing previous research that found intradisciplinary differences to be smaller than differences across disciplines (e.g., Igbaria and Baroudi 1995; Weisshaar 2017).

In a team-based learning environment, students gain valuable skills for the workplace in terms of learning to work with others and gain confidence in how to do so. Students also gain experience in evaluating peers, a task arguably distasteful yet necessary in many workplaces. Brutus, Donia, and Ronen (2013) found that students gained confidence and provided more detailed evaluations of peers with more experience in project-based learning in undergraduate business classes. More experience evaluating others in a manner that requires differentiation and articulation of qualities and behaviors that underlie choices of ratings may improve evaluation of and by both men and women in future situations requiring rating or evaluation of colleagues.

To the extent that females may benefit more, as reflected in higher peer evaluations, a cooperative learning environment of this sort might help to close the gender gap in self-promotion found by Exley and Kessler (2019). If women prefer such a learning environment and do better academically, cooperative learning in economics courses may encourage more women to major in economics or even to continue to graduate school in economics; even small changes could be significant in a discipline in which females comprise only about 37 percent of senior undergraduates nationwide and about 33 percent of first-year doctoral students (CSWEP 2018).

Notes

1. For more detail on implementing TBL in economics courses, see Clerici-Arias (2021), Green (2021), and Ruder, Maier, and Simkins (2021).

2. Histograms for peer evaluations by gender for each of the three courses are available upon request.
3. The first alternative was a gender diversity index equal to 1 minus the sum of squared proportions of each gender. A gender-balanced team would have an index equal to 0.5, while one with no gender diversity would have an index equal to 0. The second alternative was the use of two variables, each indicating the extent of male or female dominance on a team, calculated as percent male minus 0.5 for majority male teams and percent female minus 0.5 for female majority teams.
4. Indicator variables for whether or not the student being evaluated is the only female or one of two females on her team are also considered but were not found to be statistically significant, so they are not reported here.

References

Almlund, M., A. Duckworth, J. Heckman, and T. Kautz. 2011. Personality psychology and economics. In *Handbook of the economics of education*, vol. 4, ed. E. A. Hanushek, S. J. Machin, and L. Woessmann, 1–181. Amsterdam, The Netherlands: North-Holland Press.

Arbuckle, J., and B. D. Williams. 2003. Students' perceptions of expressiveness: Age and gender effects on teacher evaluations. *Sex Roles* 49 (9/10): 507–16. doi: 10.1023/A:1025832707002.

Babcock, L., and S. Laschever. 2003. *Women don't ask*. Princeton, NJ: Princeton University Press.

Bachen, C. M., M. M. McLoughlin, and S. S. Garcia. 1999. Assessing the role of gender in college students' evaluations of faculty. *Communication Education* 48 (3): 193–210. doi: 10.1080/03634529909379169.

Basow, S. A. 1995. Student evaluations of college professors: When gender matters. *Journal of Educational Psychology* 87 (4): 656–65. doi: 10.1037/0022-0663.87.4.656.

Boring, A., K. Ottoboni, and P. B. Stark. 2016. Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research* 0 (0):1–11. doi: 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1.

Brutus, S., M. B. L. Donia, and S. Ronen. 2013. Can business students learn to evaluate better? Evidence from repeated exposure to a peer-evaluation system. *Academy of Management Learning & Education* 12 (1): 18–31. doi: 10.5465/amle.2010.0204.

Castilla, E. 2008. Gender, race, and meritocracy in organizational careers. *AJS; American Journal of Sociology* 113 (6): 1479–526. doi: 10.1086/588738.

Clerici-Arias, M. 2021. Transitioning to a team-based learning principles course. *Journal of Economic Education* 52 (3): 249–56. doi: 10.1080/00220485.2021.1925184.

Committee on the Status of Women in the Economics Profession (CSWEP). 2018. *The 2018 report of the Committee on the Status of Women in the Economics Profession*. Nashville, TN: AEA, CSWEP. https://www.aeaweb.org/content/file?id=9270.

Deloitte Development LLC (Deloitte). 2016. *Global human capital trends 2016: The new organization: Different by design*. New York: Deloitte University Press.

De Paola, M., and V. Scoppa. 2015. Gender discrimination and evaluators' gender: Evidence from Italian academia. *Economica* 82 (325): 162–81. doi: 10.1111/ecca.12107.

Espey, M. 2018a. Diversity, effort, and cooperation in team-based learning. *Journal of Economic Education* 49 (1): 8–21. doi: 10.1080/00220485.2017.1397571.

——— 2018b. Enhancing critical thinking using team-based learning. *Higher Education Research & Development* 37 (1): 15–29. doi: 10.1080/07294360.2017.1344196.

Exley, C. L., and J. B. Kessler. 2019. *The gender gap in self-promotion*. Working Paper 26345. Revised May 2021. Cambridge, MA: National Bureau of Economic Research.

Falchikov, N., and J. Goldfinch. 2000. Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research* 70 (3): 287–322. doi: 10.3102/00346543070003287.

Green, A. 2021. TBL Fridays: Using team-based learning to engage in policy debates in an introductory class. *Journal of Economic Education* 52 (3): 257–63. doi: 10.1080/00220485.2021.1925186.

Igbaria, M., and J. J. Baroudi. 1995. The impact of job performance evaluation on career advancement prospects: An examination of gender differences in the IS workplace. *MIS Quarterly* 19 (1): 107–27. doi: 10.2307/249713.

Jacob, B. A. 2002. Where the boys aren't: Non-cognitive skills, returns to school and the gender gap in higher education. *Economics of Education Review* 21 (6): 589–98. doi: 10.1016/S0272-7757(01)00051-6.

Lee, H.-J., and C. Lim. 2012. Peer evaluation in blended team project-based learning: What do students find important? *Educational Technology & Society* 15 (4): 214–24.

Liu, J., D. T. Pysarchik, and W. W. Taylor. 2002. Peer review in the classroom. *BioScience* 52 (9): 824–29. doi: 10.1641/0006-3568(2002)052[0824:PRITC.2.0.CO;2]

Lyness, K. S., and M. E. Heilman. 2006. When fit is fundamental: Performance evaluations and promotions of upper-level female and male managers. *The Journal of Applied Psychology* 91 (4): 777–85. doi: 10.1037/0021-9010.91.4.777.

MacNell, L., A. Driscoll, and A. N. Hunt. 2015. What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education* 40 (4): 291–303. doi: 10.1007/s10755-014-9313-4.

Michaelsen, L. K., A. B. Knight, and L. D. Fink. 2002. *Team-based learning: A transformative use of small groups in college teaching*. Westport, CT: Praeger Publishers.

Niederle, M. 2021. Gender. In *Handbook of experimental economics*, ed. J. Kagel and A. E. Roth, vol. 2, 481–553. Princeton, NJ: Princeton University Press.

Ruder, P., M. H. Maier, and S. P. Simkins. 2021. Getting started with team-based learning (TBL): An introduction. *Journal of Economic Education* 52 (3): 220–30. doi: 10.1080/00220485.2021.1925187.

Topping, K. J., E. F. Smith, I. Swanson, and A. Elliot. 2000. Formative peer assessment of academic writing between post-graduate students. *Assessment & Evaluation in Higher Education* 25 (2): 149–66. doi: 10.1080/713611428.

Weisshaar, K. 2017. Publish and perish? An assessment of gender gaps in promotion to tenure in academia. *Social Forces* 96 (2): 529–60. doi: 10.1093/sf/sox052.

Wever, B. D., H. V. Keer, T. Schellens, and M. Valcke. 2011. Assessing collaboration in a wiki: The reliability of university students' peer assessment. *The Internet and Higher Education* 14 (4): 201–6. doi: 10.1016/j.iheduc.2011.07.003.