Summer 6-30-2018

Exploring Gender Effects on Peer Rating in Open Innovation and Crowdsourcing: A Case of Website Evaluation

Liang Chen
Department of Computer Information and Decision Management, West Texas A&M University, USA

Exploring Gender Effects on Peer Rating in Open Innovation and Crowdsourcing: A Case of Website Evaluation

Liang Chen
Department of Computer Information and Decision Management,
West Texas A&M University, USA

Abstract: Peer rating has be used by open innovation and crowdsourcing platforms to evaluate submissions and select winners because it not only represents a cheaper and more scalable way but also empowers and engages users. However, the literature on scholarly peer review suggests that peer rating may suffer from some biases. One of them is caused by gender. Therefore, this paper aims to examine gender effects on peer rating in open innovation and crowdsourcing. More specifically, we examine how judge gender and gender similarity between judge and designer affect peer rating score. This question has never been examined in the OI&C literature. Using a quasi-experimental design, we collect 1,585 evaluations and find that, overall, judge gender has no significant effect on peer rating score, but gender similarity has a negative effect. Further examinations reveal that rating mode (single-blind or double-blind) may moderate such gender effects: male judges are predicted to give a higher rating score than females when the designer's information is disclosed while in double-blind peer rating gender similarity reduces the peer rating score. This study has practical implications to the use and design of a peer rating system in OI&C platforms.

Keywords: crowdsourcing, gender effect, open innovation, peer rating, selection system

1. INTRODUCTION

As a selection system, peer rating or evaluation has been adopted by many open innovation and crowdsourcing (OI&C) platforms [1, 2]. For example, in Threadless, a crowdsourcing community of artists, all designs are created and then chosen by its community. Each week, hundreds of designed T-shirts are produced by artists and then rated by peer artists. Several other OI&C platforms such as Jovoto (for brand innovation), Zooppa (for video commercials), PimTim (for graphic design), and Lego Ideas (for Lego design) have similarly adopted peer rating as their selection method to identify best submissions in their contests. Online peer rating represents a cheaper, more scalable way of evaluating submission and selecting winners, and therefore possibly it can be a powerful new way of evaluating submission quality in OI&C platforms [3, 4, 5]. In addition, peer rating empowers and engages users, which can expand the platform and maintain users in the OI&C community [6].

Despite of many advantages, peer rating also has its bias. Similar to scholarly peer review, peer rating in OI&C relies on the evaluation from one or multiple people of similar competence to the content producer. Lee et al. [7] identify different types of bias in peer evaluation process, including bias resulting from designer and evaluator characteristics. Among those characteristics, gender is a salient one. However, the results they have reviewed are mixed: some researchers find significant gender bias but others do not. Compared with scholars in peer review, participants in OI&C communities have an even bigger gap in terms of their knowledge and experience and thus non-quality factors such as gender may affect peer rating results. In sum, gender could be an influencing factor in online peer rating. In order to address this issue, we raise our research questions: do gender effects exist in peer rating in OI&C communities? If so, are gender effects moderated by the rating mode (single- or double-blind).

Gender effects in peer rating can be driven by two factors: judge gender (i.e., do male and female judges rate the same entry differently?) and the similarity between a judge's gender and a designer's (i.e., do judges with the same gender as the designer rate the entry differently from the one with different gender?). This study focuses on the later one while controls the first one and explores the research questions in the context of web design

because it is very common task in OI&C platforms such as 99.design, Taskcn, DesignCrowd, CrowdSpring and among many others. Because there are single-blind and double-blind peer rating, we consider and compare two rating modes: the designers' information including their gender is disclosed to judges (single-blind) and not disclosed (double blind). We conducted a quasi-experimental design and collected 1,585 evaluations of websites designed by college students, made by their fellow students. We find that, overall, judge gender has no significant effect on peer rating score, but gender similarity has a negative effect. Further examinations reveal that rating mode (single-blind or double-blind) may moderate such gender effects: male judges are predicted to give a higher rating score than females when the designer's information is disclosed while in double-blind peer rating, gender similarity reduces the peer rating score. Our findings suggest that gender effects must be considered when designing peer rating system in OI&C platforms.

This remainder of this paper is organized as follows. The next section provides a theoretical background by reviewing the relevant literature and discussing gender effects. Research method is described in Section 3 and results are presented in Section 4. The final section discusses practical implications, research limitations, and future research directions.

2. BACKGROUND

2.1 Peer rating as a selection system

Selection is the process of choosing candidates from a group of potentials based on some criteria. The selection system theory identifies three ideal types of selection systems based on the "selectors": *market*, *peer*, and *expert* [1, 8]. In the market selection, consumers are selectors. In the peer selection, the selectors and the selected are part of the same group. In the expert selection, the selectors are neither producers nor consumers, but have the power to shape selection by virtual of their specialized knowledge and distinctive abilities. This literature adds that in practice a combination of selector types may occur, perhaps at different stages of the selection processes, but it is still useful to specify the dominant selectors.

As a form of peer rating, peer review has been introduced to the academia for many decades. It has provided a reliable form of scientific communication and ensured the quality of scientific research. Peer review is recognized as a required component of research validation, the academic reward system, and the scholarly publication process[7]. Peer review could be single blind, double-blind, or open.

Recently, as the Internet provides a great way for open innovation, many OI&C platforms rely on peer rating for quality control, winner selection, and reward distribution. Peer rating is recognized as an important evaluation method in innovation contests[2, 9]. Based on 33 articles and 57 real-world innovation contests, Bullinger & Moslein[2] recognize the evaluation mechanism as one of the ten key design elements for innovation contests. According to them, innovation contests can use expert evaluation, peer rating, self-assessment, or mixed method. They also find that 31 out of 57 real-world innovation contests or platforms use community functionality such as commenting functions and forums and all of them comprise any form of peer rating. Despite of its great importance and relevance to open innovation, very few studies examine this issue in the context of OI&C.

2.2 Gender impact in peer rating

Gender plays an important role in human interaction. Previous studies have examined the effects of gender and gender similarity on various decisions and outcomes such as employment interviews and recruitment, job analysis, sales, and customer service, but their results are inconsistent and mixed [10, 11]. Those studies find positive, negative, or no gender effects. Therefore, those studies, we cannot infer how gender and gender similarity may affect peer rating result in OI&C.

Gender is a salient factor influencing peer rating results. Generally, the gender effect on peer rating consists

of two components: the judge gender and the similarity between a judge's gender and a designer's. Table 1 summarizes a few studies investigating the gender impact in peer rating under various research contexts. Both Girard & Pinar[12] and Chen & Fang[13] find that female reviewees receive a higher rating from peer reviewers. Pinar & Girard[14] find that gender similarity increases peer rating results in one setting, but not in another setting.

Table 1. Research on gender effect in peer rating

Study	Main findings	Research context
[7]	This study identifies different types of bias in peer evaluation process. One of them is bias as a function of author characteristics such as nationality and gender.	Scholar publication
[12]	The gender of evaluators or presenters did not have any significant effect on presentation scores. Female students seem to be perceived as better presenters than male students.	Student presentation
[13]	Compared with male reviewers, female reviewers have a significantly higher number of high-quality reviews	Online reviews
[15]	Even though the estimates of the gender effect vary substantially across studies, men applicants have statistically significant greater odds of receiving grants than women by about 7%.	Grant application

2.3 Judge Gender

One driver of gender effect in peer rating is judge gender. The literature on marketing and psychology indicates that male and female process the same information in different ways. When judging a product, males were less oriented to visuals and more motivated extrinsically than females [16]. Similarly, Wesley et al. [17] posit that male consumers show significantly lower recreational consciousness and fashion consciousness than female consumers in their shopping activity. For example, Seock and Sauls [18] find that male and female customers have different shopping orientation and use different criteria when evaluating apparel retail stores. Coontz [19] finds significant differences between female and male judges when making their judicial decisions. However, Cooper et al. [20] find that judge gender explain little of the variation in the ratings of the job evaluation.

2.4 Gender Similarity

The similarity between a judge's gender and a designer's may also affect peer rating results. However, there are two opposite arguments about how gender similarity impacts peer rating outcomes. On the one hand, the similarity-attraction theory posits that individuals tend to like and be attracted to others who are similar, rather than dissimilar, to themselves [21]. Accordingly, gender similarity may increase peer ratings. In other words, a submission is expected to receive a higher rating score when the judge has the same gender with the designer.

On the other hand, gender dissimilarity is recognized as positive signal for social identification and will produce attraction between selectors and selectees [22]. For example, Jones et al. [23] find that customers are more likely to accept salespersons who are dissimilar to themselves. Accordingly, gender dissimilarity may yield a favorable peer rating score.

3. RESEARCH METHOD

In order to explore our research questions, we conducted a quasi-experiment design. Undergraduate students from multiple sections of a management information systems course at a major university were required to complete an individual project, which included two stages: designing a website and then evaluating websites designed by peer students. The instructor used the format provided by an open innovation platform to write a website design brief. At the first stage, students followed the brief to design a website on the same platform. At the second stage, students were randomly assigned to be a judge to evaluate 10-13 websites design by students

from a different section. The data was collected at the website evaluation stage. In total, 139 students participated in this project. Among them, 73 (52.5%) were male while the remaining were female. One website was dropped because of no accessibility. In total, 1585 evaluations were obtained from 138 websites and 139 judges.

Our dependent variable is overall peer rating score of each website. It ranges from 1 (very poor) to 5 (excellent). Among 1585 evaluations, 1454 include comments. There were 11,593 words in all these comments and, on average, each comment has 7.97 words. We use text mining techniques and generate a sentiment score based on judges' comments. In sentiment analysis, each comment receives a score describing it to be either positive or negative. This was conducted in RapidMiner 7.6. The sentiment score from each evaluation is considered as a supplementary dependent variable to check the reliability of our main dependent variable. Some sample comments and their sentiment score are shown in Table 2.

Table 2. Sample comments and their sentiment score

Comments	Sentiment Score
The photo gallery pics were not organized	-0.63
It's nearly the same with other websites	-0.36
This design was a little confusing for me	-0.33
The design of this website was quite boring	-0.25
I thought the website had some cool information, but the design of it was confusing, as I was not very sure what to look for.	-0.04
Very thorough website. It seems like you're involved in a lot. Maybe include some pictures to spice up the page a little bit.	0.12
Purpose isn't explicitly stated. Design is very well structured and easy on the eyes. Very thorough effort.	0.37
Good concept, purpose was a bit unclear	0.38
Loved the colors and design of this website!	0.50
This site was well put together and contained a lot of thorough information.	0.51
Author of website is enthusiastic!	0.63

Our independent variable is gender similarity (1: the judge and designer have the same gender; 0: the judge and designer have different genders). We control judge gender (1: male; 0: female), and website attributes, including purpose (how successfully this website serves a clear purpose such as introducing a person, a company, or a place or conducting business activities such as shopping or financial service), design (how successfully this website is well-designed such as well-organized content, appealing visuals, easy navigation), and originality (how successfully this website is distinguishable from other websites and gives you something that you cannot find elsewhere). These three attributes are usually considered as the top criteria to evaluate a website. Each of the three variables ranges from 1 (very unsuccessfully) to 5 (very successfully).

In addition, designers were allowed to voluntarily choose whether or not disclose their gender information such as biography and pictures on their websites. Doing so, we created an additional factor, rating mode, to indicate whether the evaluation is single-blind (i.e., the judge can view the designer's gender while the designer does not the judge's gender) or double-blind. We will test whether rating mode moderate gender effects on peer rating.

4. RESULTS

The mean and standard deviation of each variable and correlation coefficients among all the variables are

presented in Table 3. Each evaluation represents a designer-judge pair. About a half of 1585 evaluations are evaluated by male judges and 52% of designer-judge pairs have the same gender. Peer rating score has a significantly positive relationship with Sentiment score, indicating a good reliability of our main dependent variable.

We first run a linear regression model with the peer rating score as the dependent variable for all the sample. We find that gender similarity has a significant but marginal negative influence on peer rating score. We do not find a significant relationship between judge gender and peer rating score. In addition, the positive and significant regression coefficients of all the three controlled variables indicate that the three website attributes purpose, design, and originality play an important role in determining peer rating score.

Table 3. Descriptive statistics of seven variables

Variable	Mean	Std. Dev.	1	2	3	4	5	6	7
1. Judge gender	0.52	0.500	1.00						
2. Gender similarity	0.52	0.500	0.06	1.00					
3. Purpose	4.64	0.745	-0.11	-0.02	1.00				
4. Design	4.29	0.937	-0.09	-0.03	0.59	1.00			
5. Originality	4.43	0.848	-0.15	-0.05	0.52	0.58	1.00		
6. Peer rating score	4.47	0.809	-0.11	-0.06	0.75	0.83	0.74	1.00	
7. Sentiment score	0.20	0.222	0.00	0.00	0.27	0.31	0.23	0.34	1.00

Note: Sample size is 1585 for the first six variables and 1454 for the last variable, Sentiment score.

Table 4. Linear regression model with overall rating score as the dependent variable

	All sample	Single-blind	Double-blind
	Coefficient. (Std. Error)	Coefficient. (Std. Error)	Coefficient. (Std. Error)
Independent Variable			
Gender Similarity	- 0.050 (0.0158) **	- 0.042 (0.0333)	- 0.040 (0.0186) *
Controlled variables			
Judge Gender	0.016 (0.0160)	0.067 (0.0335) *	0.008 (0.0188)
Purpose	0.348 (0.014) ***	0.391 (0.025) ***	0.330 (0.017) ***
Design	0.397 (0.011) ***	0.358 (0.024) ***	0.409 (0.013) ***
Originality	0.289 (0.012) ***	0.309 (0.023) ***	0.287 (0.014) ***
Sample Size	1585	399	1150
Adjusted R Square	0.848	0.860	0.846

* p<0.05, ** p<0.01, *** p<0.001.

In order to check whether rating mode (single-blind or double-blind) moderate the effects of judge gender and gender similarity on peer rating score, we divide the whole sample into two groups and run the same model individually for each group. Interestingly, gender effects are moderated by rating mode. Specifically, when the designer information is not disclosed (double-blind rating), gender similarity still has a negative effect on peer rating score, which is consistent with the results derived from the whole sample. A potential reason is that a judge may get some cues from the web design elements such as the use of color or pictures to determine the designer's gender. This explanation is found from the research on scholarly reviews, which reveals that even though author anonymity can prevent reviewer bias in the double-blind review, reviewers can often identify the

author through their writing style, subject matter, or self-citation [7]. In our case, judges may identify the designer' gender based on their wording, colors, pictures, interests, or hobbies on their website even though they may not identify the designer.

When the designer information is disclosed to judges, the effect of gender similarity is insignificant, but the judge gender has a significant and positive effect on peer rating score, indicating that male judges give a higher rating to the designer no matter whether they have the same gender or not. Under the double-blind rating mode, the design of a website makes the largest contribution to the overall peer rating score, while under the single-blind rating mode, the purpose of a website is the greatest contributor. These comparisons suggest the existence of a potential moderation effect of rating mode (single-blind or double-blind), which can be further examined in future research.

When the values of controlled variables are fixed, Figure 1 presents the difference of designers' rating scores under two rating modes. Interestingly, no matter whether the designer's gender information is disclosed or not, female designers always receive the highest rating score from male judges, but the lowest rating from female judges. No matter whether the designer's gender information is disclosed or not, female designers always receive a higher rating from male judges than that from female judges. When the designer's information is disclosed, male designers receive a higher rating score from male judges than that from female judges. However, when the designer's gender information is not disclosed, male designers receive a higher rating from female judges than from the male judges. This finding indicates that the gender effect on male designers' rating is moderated by rating mode.

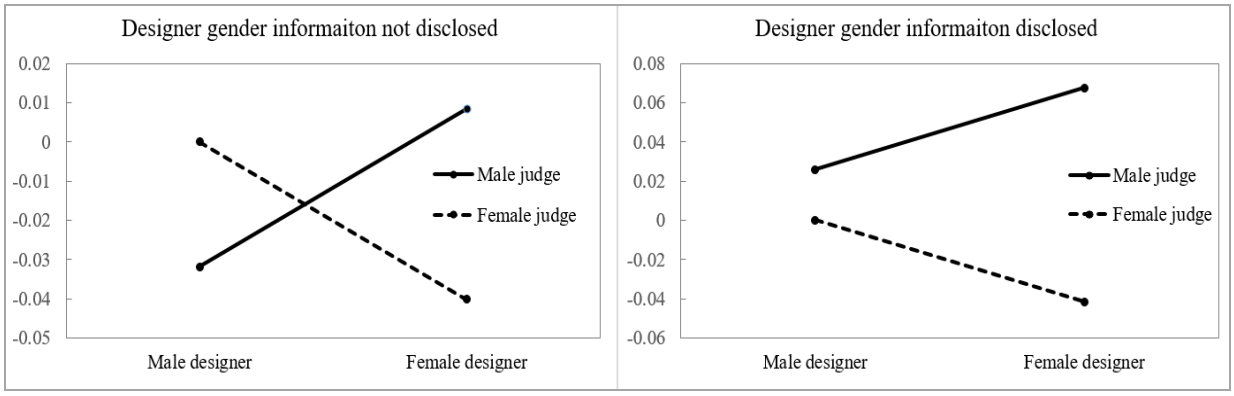

Figure 1. The difference of designers' rating scores under two rating modes

5. CONCLUSIONS AND DISCUSSIONS

This paper aims to investigate gender effects on peer rating in OI&C: whether judge's gender and the gender similarity between judge and designer may influence peer rating in the case of website design. Overall, we find that judge gender does not affect the peer rating score, but gender similarity decreases the peer rating score. Our case supports the negative impact of gender similarity, which may indicate that dissimilarity can attract higher peer ratings in OI&C. In addition, we find that the effects of judge gender and gender similarity on peer rating score seem to be moderated by rating mode (single-blind or double-blind): in single-blind peer rating, male judges are likely to give a higher rating score than females while in double-blind peer rating, gender similarity reduces the peer rating score.

This study has a few practical implications. Because gender effects do exist in the peer rating, no matter whether the rating is single-blind or double-blind, OI&C platforms should find a good way to avoid or minimize the gender effect. For example, they can keep a balance of male and female judges in each open innovation project. Compared with the overall rating score (main dependent variable), sentiment score (supplementary

dependent variable) is less sensitive to gender effect. Therefore, when evaluating an OI&C project, both objective rating based on evaluation rubric and subjective comments should be considered.

This study further confirms the significant and dominant role of website's purpose, design, and originality in website rating. A well-designed and creative website with a clear purpose always deserve a higher rating from peer judges. However, when all these factors are the same, gender effects cannot be ignored because even a marginal increment in peer rating can change the ranking of competitors in an OI&C contest.

This study also has a few limitations. First of all, we did not control the disclosure decision of designer's information in their submissions. Future experiment can add this treatment to systematically examine the effect of rating mode. Secondly, in our quasi-experiment, judges were randomly assigned to evaluate all 11-13 websites systematically, but in reality judges are voluntary to evaluate particular submissions in a non-systematic manner. Thirdly, most participants in this study are naïve, rather than professional designers, so it is possible that they disclosed their information unintentionally. All the limitations can be addressed in the future by designing a more strict experiment with rating mode as a factor, including other types of tasks such as logo design, or incorporating the real rating behavior data from an open innovation platform.

REFERENCES

[1] Gemser, G., Leenders, M.A., and Wijnberg, N.M. (2008).Why Some Awards Are More Effective Signals of Quality Than Others: A Study of Movie Awards†. Journal of Management, 34(1): 25-54.

[2] Bullinger, A.C., and Moeslein, K. (2011).Innovation Contests: Systematization of the Field and Future Research. International Journal of Virtual Communities and Social Networking, 3(1): 1-12.

[3] Howe, J. (2008). Crowdsourcing: How the power of the crowd is driving the future of business. Random House.

[4] Brabham, D.C. (2009).Crowdsourcing the public participation process for planning projects. Planning Theory, 8(3): 242-262.

[5] Majchrzak, A., and Malhotra, A. (2013).Towards an information systems perspective and research agenda on crowdsourcing for innovation. Journal of Strategic Information Systems, 22(4): 257-268.

[6] Semmann, M., and Grotherr, C. (2017). How to Empower Users for Co-Creation-Conceptualizing an Engagement Platform for Benefits Realization. *The 13th International Conference on Wirtschaftsinformatik* St.Gallen, Switzerland: AIS.

[7] Lee, C.J., Sugimoto, C.R., Zhang, G., and Cronin, B. (2013).Bias in peer review. Journal of the Association for Information Science and Technology, 64(1): 2-17.

[8] Wijnberg, N.M. (2011).Classification systems and selection systems: The risks of radical innovation and category spanning. Scandinavian Journal of Management, 27(3): 297-306.

[9] Adamczyk, S., Bullinger, A.C., and Möslein, K.M. (2012).Innovation Contests: A Review, Classification and Outlook. Creativity and Innovation Management, 21(4): 335-360.

[10] Graves, L.M., and Powell, G.N. (1995).The Effect of Sex Similarity on Recruiters' evaluations Of Actual Applicants: A Test of the Similarity‐Attraction Paradigm. Personnel Psychology, 48(1): 85-98.

[11] Hardin, J.R., Reding, K.F., and Stocks, M.H. (2002).The effect of gender on the recruitment of entry-level accountants. Journal of Managerial Issues): 251-266.

[12] Girard, T., and Pinar, M. (2009).An exploratory study of gender effect on student presentation evaluations: Does gender similarity make a difference? International Journal of Educational Management, 23(3): 237-251.

[13] Chen, L., and Fang, J. (2017). Users' Attributes Associated with High-Quality Review in Online Communities. *WHICEB 2017 Proceedings*, Wuhan, China: AIS.

[14] Pinar, M., and Girard, T. (2006). Student Perceptions of Class Presentations: Does Gender Impact The Evaluations? , *Marketing Management Association 2006 Educators' Conference Proceedings*, pp. 28.

[15] Bornmann, L., Mutz, R., and Daniel, H.-D. (2007).Gender differences in grant peer review: A meta-analysis. Journal of Informetrics, 1(3): 226-238.

[16] Holbrook, M.B. (1986).Aims, concepts, and methods for the representation of individual differences in esthetic responses to design features. Journal of consumer research, 13(3): 337-347.

[17] Wesley, S., LeHew, M., and Woodside, A.G. (2006).Consumer decision-making styles and mall shopping behavior: Building theory using exploratory data analysis and the comparative method. Journal of Business Research, 59(5): 535-548.

[18] Seock, Y.-K., and Sauls, N. (2008).Hispanic consumers' shopping orientation and apparel retail store evaluation criteria: An analysis of age and gender differences. Journal of Fashion Marketing and Management: An International Journal, 12(4): 469-486.

[19] Coontz, P. (2000).Gender and judicial decisions: Do female judges decide cases differently than male judges? Gender issues, 18(4): 59-73.

[20] Cooper, E.A., Doverspike, D., Barrett, G.V., and Alexander, R.A. (1987).Sex bias in job evaluation: The effect of sex on judgments of factor and level weights. Educational and psychological measurement, 47(2): 369-375.

[21] Byrne, D.E. (1971). The Attraction Paradigm. New York: Academic Press.

[22] Schaffer, B. (2016). Dissimilarity-Attraction in Teams: New Ideas for Workplace Diversity Research. *Academy of Management Proceedings*: Academy of Management, pp. 12344.

[23] Jones, E., Moore, J.N., Stanaland, A.J., and Wyatt, R.A. (1998).Salesperson race and gender and the access and legitimacy paradigm: Does difference make a difference? Journal of Personal Selling & Sales Management, 18(4): 71-88.