

# An Experimental Comparison between Self- and Third-Party Evaluations

Takehisa Kumakawa

Faculty of Business Administration, Osaka University of Economics, Osaka, Japan

Email: [kumakawa@osaka-ue.ac.jp](mailto:kumakawa@osaka-ue.ac.jp)

Received 27 April 2015; accepted 18 July 2015; published 21 July 2015

Copyright © 2015 by author and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

---

## Abstract

How to pick up the true meaning of messages exchanged in the laboratory is an important issue for experimental research. The present study investigates, by experimentally comparing self- and third-party evaluations, to what extent self-evaluations by message receivers can be relied on. After standard public-good game, subjects receive a free-form written message evaluating their decision and self-evaluate its content from their counterparts. Third-party evaluators also evaluate the content independently. A comparison between both evaluations shows that a significant proportion of them agree. Firm evidence of a self-serving bias cannot be found.

## Keywords

Self-Evaluation, Third-Party Evaluation, Self-Serving Bias

---

*Few people are wise enough to prefer useful criticism to the sort of praise which is their undoing.*

—La Rochefoucauld, *Maxims*

## 1. Introduction

Non-restricted communication plays an important role in economic decision-making. Experimental literature continues to accumulate evidence showing that the behavior of subjects significantly differs in depending on whether they are allowed to send messages to each other or not (e.g., Cooper and Kagel [1]; Charness and Dufwenberg [2]; Schotter and Sopher [3]; Kimbrough *et al.* [4]; Sutter and Strassmair [5]). That is, cheap talk is not so cheap after all. One important issue here is how to pick up the true meaning of messages exchanged in the laboratory. Message receivers correct their behavior according to the content of a message, and message senders expect in advance that this will happen and write the message. Therefore, to understand why such a message has a significant impact, researchers have to reproduce the true interpretation of a message by the message receiver.

Message receivers' self-reports look sound, but they may produce a severe self-serving bias (e.g., Babcock *et al.* [6]; Babcock and Loewenstein [7]; Offerman [8]); *i.e.*, receivers may interpret a message in their own favor

for some reasons when are asked to do so. For example, they may want to justify their behavior to preserve their self-image. Alternatively, they may want to appear to be nice to the experimenter. By contrast, third-party evaluations may be somewhat more objective but differ from the message receiver's interpretation being made during the experiment.

However, very little is known to what extent subjects' self-evaluations and third-party evaluations agree or disagree. This study provides evidence on this question. Using a message exchanges an experiment described below, self-evaluation of a message received by the subjects is compared with third-party evaluations of it. A significant proportion of both evaluations agree, suggesting that the subjects' self-evaluations are, at least to some extent, reliable. Firm evidence of a self-serving bias cannot be found.

## 2. Experimental Design

### 2.1. Message Exchange Experiment

The message exchange experiment, which comprises two stages, is described as follows. In the first stage, paired subjects play a standard public-good game; they simultaneously decide how much they would invest in the public good from 20 units of an endowment. A zero-investment is the dominant strategy, while investing the whole endowment leads to a Pareto-efficient allocation (see the appendix for a detailed description of the voluntary contribution mechanism used in the experiment). In the second stage, after the decision of their partners has been revealed, subjects write a free-form message evaluating their partners' contribution and send it to them. The message is inputted via a keyboard, and is not handwritten. After the subjects have sent their message, it is displayed on their partners' computer screens. After confirming the content of the message, the subjects classify it into the following three evaluation indexes: positive, neutral, or negative.

The experiment was conducted at Osaka University. Twenty subjects participated in each of two sessions<sup>1</sup>. The experiment required approximately one hour, and the average payoff per subject was \$23.61.

### 2.2. Third-Party Evaluations of Messages

After the message exchange experiment, an additional 12 students were employed as third-party evaluators. After a detailed description of the message exchange experiment, they simultaneously and independently classify the messages actually written in the experiment according to their content into the same three evaluation indexes. Among these 12 evaluators' decisions on each message, the most popular one was adopted as the third-party evaluation<sup>2</sup>.

Comparing self- and third-party evaluations of a message, we define a self-serving bias as follows.

**Definition 1.** *We say there exists a self-serving bias if (i) the self-evaluation of a message is neutral or positive but the third-party evaluation of it is negative or (ii) the self-evaluation of a message is positive but the third-party evaluation of it is neutral.*

In other words, if the receiver of a message interprets it more positively than third-party evaluators do, a self-serving bias is indicated. Of course, we can consider the opposite bias, such as a sort of self-discipline bias<sup>3</sup>.

## 3. Results

### 3.1. Comparison between Self- and Third-Party Evaluations

**Figure 1** shows the linkage between subjects' relative contributions compared to their partners' and the types of messages they received and rated. The horizontal axis denotes the relative difference between the contributions of paired subjects, and the vertical axis denotes the frequency of each type of message. For example, three subjects contributed 15 units more than their counterparts did and received a positive message. As seen in **Figure 1**, subjects who contributed more than their counterparts did tend to receive a positive message, and those who contribute less, a negative message. However, the relationship is not simply linear<sup>4</sup>.

This nonlinear relationship seemingly implies that the messages are distorted by their receivers, but this is not

<sup>1</sup>The backgrounds of the sampled students were diverse and included fields other than economics.

<sup>2</sup>In the event of a tie, the investigator would have had a casting vote; however, there were no ties.

<sup>3</sup>The definition of a self-discipline bias is as follows: we say there exists a self-discipline bias if (i) the self-evaluation of a message is neutral or negative but the third-party evaluation of it is positive or (ii) the self-evaluation of a message is negative but the third-party evaluation of it is neutral.

<sup>4</sup>No linear correlation between the relative contribution differences and the frequency of each type of message was statistically confirmed.

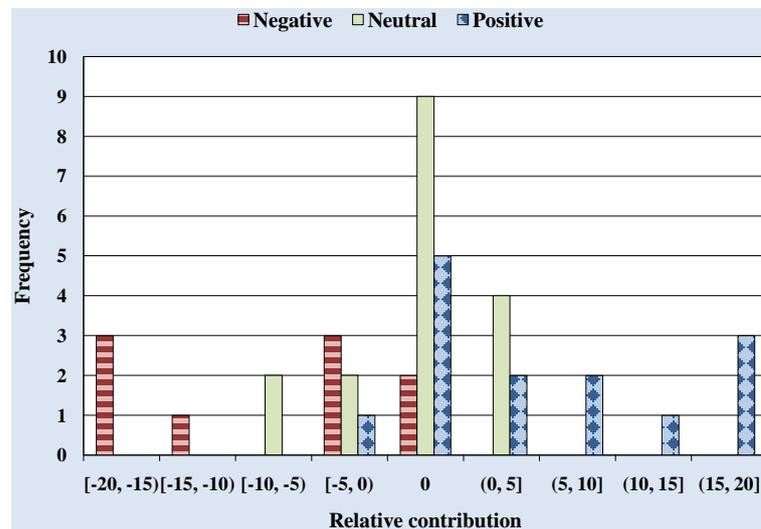


Figure 1. Relative contributions and the types of messages.

true. **Table 1** compares the evaluations of messages by the subjects themselves and those by third-party evaluators. The data in parentheses, those with underlines, and those in brackets represent self-serving evaluations, agreed evaluations, and self-discipline evaluations, respectively. Self-serving and self-discipline evaluations account for 17.5% and 7.5%, respectively, and the remaining 75% are agreed evaluations. At least the experimental data show that a significant proportion of self-evaluations are reliable in the sense that they match third-party evaluations. Firm evidence of a self-serving bias (and that of the opposite bias) could not be found.

### 3.2. Does Majority Opinion Matter?

In addition to the direct comparison of self- and third-party evaluations presented above, a further investigation was conducted on the agreement between both evaluations. In the analysis so far, the most popular evaluation among third-party evaluators was adopted as the average opinion and the number of votes in its favor was neglected. Here, we investigate the relationship between the number of votes for the most popular evaluation and the probability that both evaluations agree. Specifically, the following probit model was estimated:

$$d_A = \beta_0 + \beta_1 n + \beta_2 \text{Max}\{0, c_i - c_j\} + \beta_3 \text{Max}\{0, c_j - c_i\}$$

where the dependent variable is a dummy, which equals 1 if subject  $i$ 's self-evaluation of a message received from subject  $j$  agrees with third-party evaluations and 0 otherwise. The independent variables are the number of votes for the most popular evaluation of the message by third-party evaluators and the absolute positive and negative differences between the contributions of subjects  $i$  and  $j$ .

The results are summarized in **Table 2**. First, as is apparent from **Table 2**, the number of votes for the most popular evaluation does not have a significant effect on the agreement between self- and third-party evaluations ( $p = 0.780$ ). Intuitively, the more difficult a message is for third-party evaluators to judge, the more likely it is that the receiver's evaluation would be influenced by a self-serving bias. However, this possibility was ruled out. Considering Result 1, self-evaluations accord well with third-party evaluations, even for a message on which opinions are divided among third-party evaluators.

Second, although the absolute positive and negative differences between contributions were not significant at the 10% level ( $p = 0.196$  and  $0.206$ , respectively), the estimated coefficients for both were small positive values. Some subjects and third-party evaluators might use this information, in addition to the content of a message, to evaluate it.

## 4. Conclusions

The present study provides the evidence on the extent to which self-evaluation of a message received by subjects can be relied on. The experimental data confirm that (i) their self-serving bias is not large and (ii) their self-

**Table 1.** Self- and third-party evaluations of messages.

		Third-party evaluation		
		Positive	Neutral	Negative
Self-evaluation	Positive	8	(6)	(0)
	Neutral	[2]	14	(1)
	Negative	[1]	[0]	8

Note: The data in parentheses, those with underlines, and those in brackets represent self-serving evaluations, agreed evaluations, and self-discipline evaluations, respectively.

**Table 2.** Results of the probit regression.

Independent variable	Coefficient
Constant	-0.122 (0.937)
Number of votes	0.042 (0.780)
Absolute positive difference	0.118 (0.196)
Absolute negative difference	0.111 (0.206)

Notes: The dependent variable equals 1 if a subject's self-evaluation agrees with the third-party evaluation, and 0 otherwise. The numbers in parentheses represent *p*-values.

evaluations accord well with third-party evaluations, even when a message is relatively difficult for third-party evaluators to judge. A positive interpretation of these results will imply that experimental researchers can consider subjects' self-evaluations during an experiment as reliable data for analytical purposes, at least to some extent.

However, even though the number of observations is small, some subjects interpret a message more positively—but not more negatively—than third-party evaluators do. When and how often cheating occurs is still an open question, left for future research.

Finally, a limitation of this study should be mentioned. The discussion thus far implicitly assumes that third-party evaluators will evaluate messages objectively and neutrally, at least to some extent. However, if taking their psychological factors as human beings into account, this may not always be true. For example, many studies point out that people sometimes behave spitefully, that is, their behavior intends to make others suffer (monetary or nonmonetary) losses (e.g., Jensen [9]; Leibbrandt and López-Pérez [10]). By doing so, they can derive pleasure from others' misfortune. This behavior is believed to be due to aversion to inequity and/or competitive spirit. If a third-party evaluator has such a spiteful preference, he/she may feel jealous of a player who has received praise for his/her good behavior, and consciously give him/her a low evaluation<sup>5</sup>. To verify the neutrality of third-party evaluations and its robustness, a systematic investigation is needed.

## Acknowledgements

Special thanks were due to Yuki Hamada and Keiko Takaoka, who helped conduct the experiment. This research was supported by Grants-in-Aid for JSPS Fellows 211071 and 231657 from the Japan Society for the Promotion of Science.

<sup>5</sup>In the actual experiment, for example, a subject who had contributed unilaterally to the public good got the following message: "You are the best of good fellows."

## References

- [1] Cooper, D.J. and Kagel, J.H. (2005) Are Two Heads Better Than One? Team versus Individual Play in Signaling Games. *American Economic Review*, **95**, 477-509. <http://dx.doi.org/10.1257/0002828054201431>
- [2] Charness, G. and Dufwenberg, M. (2006) Promises and Partnership. *Econometrica*, **74**, 1579-1601. <http://dx.doi.org/10.1111/j.1468-0262.2006.00719.x>
- [3] Schotter, A. and Sopher, B. (2007) Advice and Behavior in Intergenerational Ultimatum Games: An Experimental Approach. *Games and Economic Behavior*, **58**, 3650-393. <http://dx.doi.org/10.1016/j.geb.2006.03.005>
- [4] Kimbrough, E.O., Smith, V.L. and Wilson, B.J. (2008) Historical Property Rights, Sociality, and the Emergence of Impersonal Exchange in Long-Distance Trade. *American Economic Review*, **98**, 1009-1039. <http://dx.doi.org/10.1257/aer.98.3.1009>
- [5] Sutter, M. and Strassmair, C. (2009) Communication, Cooperation and Collusion in Team Tournaments—An Experimental Study. *Games and Economic Behavior*, **66**, 506-525. <http://dx.doi.org/10.1016/j.geb.2008.02.014>
- [6] Babcock, L., Loewenstein, G., Issacharoff, S. and Camerer, C. (1995) Biased Judgments of Fairness in Bargaining. *American Economic Review*, **85**, 1337-1343.
- [7] Babcock, L. and Loewenstein, G. (1997) Explaining Bargaining Impasse: The Role of Self-Serving Biases. *Journal of Economic Perspectives*, **11**, 109-126. <http://dx.doi.org/10.1257/jep.11.1.109>
- [8] Offerman, T. (2002) Hurting Hurts More Than Helping Helps. *European Economic Review*, **46**, 1423-1437. [http://dx.doi.org/10.1016/S0014-2921\(01\)00176-3](http://dx.doi.org/10.1016/S0014-2921(01)00176-3)
- [9] Jensen, K. (2010) Punishment and Spite, the Dark Side of Cooperation. *Philosophical Transactions of the Royal Society B*, **365**, 2635-2650. <http://dx.doi.org/10.1098/rstb.2010.0146>
- [10] Leibbrandt, A. and López-Pérez, R. (2011) The Dark Side of Altruistic Third-Party Punishment. *Journal of Conflict Resolution*, **55**, 761-784. <http://dx.doi.org/10.1177/0022002711408010>

## Appendix: The Voluntary Contribution Mechanism

The experiment uses the following voluntary contribution mechanism. There are two subjects,  $a$  and  $b$ , with subject  $i$  ( $=a, b$ ) having  $w_i$  units of an endowment of a private good. Each subject faces a decision regarding splitting  $w_i$  between his or her own consumption of the private good ( $x_i$ ) and investment ( $y_i$ ) in the public good ( $y$ ). From the investment, each subject enjoys  $y = y_a + y_b$ ; that is, the level of the public good is the sum of the investments made by the two subjects. Therefore, each subject's decision problem is to maximize his or her own payoff  $u_i(x_i, y)$ , subject to  $x_i + y_i = w_i$ . All subjects have the same payoff function, specified as follows:

$$u_i(x_i, y) = x_i + ay$$

where  $(w_a, w_b) = (20, 20)$  and  $a = 0.7$ , the latter of which is the marginal per-capita return from an investment in the public good.

Within these parameters, making no investment in the public good (*i.e.*, complete free-riding) is the dominant strategy for each subject in the one-shot game. Accordingly, the level of the public good is 0 in the dominant-strategy equilibrium. By contrast, the aggregate payoff of the two subjects is maximized when each subject invests all 20 units of his or her endowment (*i.e.*, full cooperation).