Comment on Camerer et al. (2016) "Evaluating replicability of laboratory experiments in economics."

We applaud the effort of Camerer et al. (2016) to replicate studies in experimental economics. We were pleased to see that the results from the replication study strongly support the findings from Ericson and Fuster (2011). However, we believe the classification of our study as "not replicated" is misleading because the authors focus on one particular statistical test from our paper that is both less powerful and more restrictive than regression specifications reported in our paper and in Camerer et al.'s replication report. These regressions strongly replicate the results from our original paper, both in terms of statistical significance and in terms of magnitudes of the effect.

Camerer et al. (2016)'s report* of Ericson and Fuster shows that the effect studied is significant at $p<0.01$ in the paper's preferred regression specifications (see Table 2, columns (2) to (5)) and the estimated effect size is almost identical. (Similarly, the simple comparisons of willingness to accept (WTA) for a mug, and also of log(WTA) across treatments are significant at $p<0.01$). However, for classification purposes, Camerer et al. focus on a test that is less powerful than the regression and imposes undesirable functional form restrictions: the t-test for whether the difference in log(WTA) for a mug versus log(WTA) for a pen is the same in the low versus high expectations conditions. This t-test gives a p-value of 0.055, and hence counts as "not replicated" relative to the $p=0.05$ threshold. This is unfortunate, since we made it clear, in both Ericson and Fuster (2011) and our correspondence with the authors prior to the replication study, that the regression-based results should be preferred not only on power grounds but also due to the more flexible functional form.

While the authors have noted that they needed to pick a single test to focus on for the purposes of their prediction market, we believe (both now and in advance) they picked the theoretically incorrect, less powerful test, and thus led our study to be misleadingly classified as "not having replicated".

We are grateful for this project and its contributions to scientific knowledge, and want to clarify the record.

Keith Marzilli Ericson and Andreas Fuster

*Report available at:
http://experimentaleconreplications.com/finalreports/Ericson%20&%20Fuster%202011.pdf

References:

Camerer, C. et al. Evaluating replicability of laboratory experiments in economics. *Science.* 10.1126/science.aaf0918 (2016).

Ericson, K and A Fuster. Expectations as Endowments: Expectations as endowments: Evidence on reference-dependent preferences from exchange and valuation experiments. *Q. J. Econ.* **126**, 1879–1907 (2011). doi:10.1093/qje/qjr034

Excerpt from Camerer et al. (2016)'s report:

Table 2: Original results and replication results for regression (1) to (5) in Table II in the original study. The dependent variable is $ln(WTA_{mug})$.

	Original Study					*Replication Study*				
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
Treatment MH	0.194	0.266**	0.306**	0.308**	0.290**	0.358***	0.266***	0.269**	0.278***	0.292***
	(0.143)	(0.124)	(0.128)	(0.126)	(0.130)	(0.128)	(0.104)	(0.106)	(0.107)	(0.108)