Institute for Empirical Research in Economics
University of Zurich

Working Paper Series
ISSN 1424-0459

Working Paper No. 437

A Correction and Re-Examination of "Stationary Concepts for Experimental 2 x 2 Games"

Christoph Brunner, Colin F. Camerer, and Jacob K. Goeree

September 2009

A Correction and Re-Examination of "Stationary Concepts for Experimental 2×2 Games"

Christoph Brunner, Colin F. Camerer, and Jacob K. Goeree*

September 15, 2009

Abstract

Selten and Chmura (*American Economic Review*, June 2008, 98(3), 938-966) recently reported experimental laboratory results for 2×2 games with unique mixed-strategy equilibria used to compare Nash equilibrium with four other stationary concepts: quantal response equilibrium, action-sampling equilibrium, payoff-sampling equilibrium, and impulse balance equilibrium. They conclude that impulse balance equilibrium performs best, and, in particular, significantly outperforms quantal response equilibrium. We reanalyze their data and correct some errors. The reanalysis shows that Nash clearly fits worst but the four other concepts perform about equally well. It is surprising that four models, which are so conceptually different, are so close in accuracy, and following Selten and Chmura's suggestion, we report new analysis of previous experiments on 2×2 games with unique mixed-strategy equilibria. These additional tests show the importance of the loss aversion that is hardwired into impulse balance equilibrium: when the other non-Nash stationary concepts are augmented with loss aversion they outperform impulse balance equilibrium.

*Brunner and Camerer: Division of the Humanities and Social Sciences, California Institute of Technology, Mail code 228-77, Pasadena, CA 91125, USA. Goeree: Institute for Empirical Research in Economics, University of Zürich, Blümlisalpstrasse 10, CH-8006, Zürich, Switzerland, and Division of the Humanities and Social Sciences, California Institute of Technology, Mail code 228-77, Pasadena, CA 91125, USA. We would like to thank participants at the Economic Science Association meetings in Tucson (November, 2008) for valuable feedback. We gratefully acknowledge financial support from the National Science Foundation (SES 0551014), the Gordon and Betty Moore Foundation, and the Dutch National Science Foundation (VICI 453.03.606).

1. Introduction

A recent paper by Reinhard Selten and Thorsten Chmura (2008) (henceforth SC) reports laboratory results for 12 different 2×2 games with a unique mixed-strategy equilibrium. These binary-choice games are relatively simple and provide a natural testbed for alternative models that aim to predict long-run, or stationary, outcomes of play. SC consider five such models: Nash equilibrium, quantal response equilibrium, action-sampling equilibrium, payoff-sampling equilibrium, and impulse balance equilibrium.

Nash equilibrium subsumes that player have correct beliefs about others' play and that players best respond to those beliefs. Quantal response equilibrium (QRE) replaces the requirement of best responses with "better responses," i.e. players are *more likely* to choose the option with the higher expected payoff but they don't necessarily choose the best option all the time. QRE does assume that players' beliefs are correct on average, i.e. beliefs are not systematically biased. Action sampling equilibrium describes the long-run outcome when players best respond to a finite sample of their opponent's previous actions.[1] Payoff sampling equilibrium describes the long-run outcome when players form two finite samples of their past payoffs, one for each option, and select the one with the highest total payoff. Finally, impulse balance equilibrium is based on the idea that players take into account foregone payoffs. If the option not chosen would have yielded higher payoffs, then there is an "impulse" to change (and, importantly, 'losses' of foregone payoff are weighted twice as heavily as gains). Impulse balance equilibrium corresponds to the long run outcome where, for both players, expected impulses are equal across the two options.

SC conclude that Nash and QRE fit worse than the other three concepts. They write:

> It is remarkable that the newer concepts of impulse balance equilibrium, payoff sampling equilibrium, and action-sampling equilibrium clearly outperform the more established concepts of quantal response equilibrium [QRE] and Nash equilibrium. All the relevant comparisons are highly significant. This is perhaps the most important result of the statistical tests. (p. 962)

[1] Action sampling equilibrium is closely related to the "stochastic learning equilibrium" concept introduced by Jacob K. Goeree and Charles A. Holt (2002) where players make a noisy best response to a weighted average of their opponents' past decisions. Rather than putting a weight of 1 on a fixed number of past observations and a weight of 0 on observations that are in the more distant past, as in action sampling equilibrium, the stochastic learning equilibrium assumes that weights decline continuously for more distant observations (e.g. geometrically). Stochastic learning equilibrium has been shown to yield an improved fit over QRE in some contexts (see, e.g., C. Monica Capra, Goeree, Rosario Gomez, and Holt, 2002).

The first point of this comment is that the model fits for two of the five concepts – QRE and action-sampling – are incorrect for all 12 games.[2] We report the correct results for these two models (and some other small corrections). The corrected fits for QRE are close to the other three non-Nash concepts, which overturns the most novel (and to some, surprising) part of their conclusion – viz., QRE fits as well as the other concepts, not worse (as SC concluded).[3]

Fit measures and statistical tests show that the four non-Nash models are about equally accurate. SC note this fact (but for three models, not all four) and suggest a research direction as follows:

> It is not easy to understand why the predictions of the three newer concepts are not very far apart, in spite of the fact that they are based on very different principles. This is perhaps peculiar to our sample. *It would be desirable to devise experiments that permit a better discrimination among the three concepts.* (p. 965, emphasis ours).

The second point of the comment is to extend the scope of their comparative analysis, by showing how two different games reported several years ago *do* "permit a better discrimination" among some of the concepts. The first game was explicitly designed to show that *no* quantal response equilibrium (logit or otherwise) could explain observed behavior (see Game 4 and Proposition 1 in Goeree, Holt, and Thomas R. Palfrey, 2003). Applying impulse balance equilibrium to this game works like 'magic:' it explains observed behavior almost perfectly. So this game is capable of differentiating between two of the concepts – impulse balance equilibrium and risk-neutral QRE – that fit equally well in SC's data.[4]

The results also highlight one of the crucial assumptions underlying impulse balance equilibrium: impulses are defined relative to a security level (the max-min payoff) and it is assumed that losses with respect to this security level are weighed *twice as much* as gains. While impulse balance equilibrium is ostensibly a parameter-free concept (since the loss aversion coefficient is fixed to 2), this additional assumption about players' different

[2] A referee also asked us to correct a typo on page 945 of the SC paper in paragraph 4; "row R" should be "column R".

[3] It is true that with the corrected analysis, Nash predictions *do* fit worse than the other four concepts. However, the ability of other models to explain deviations from Nash play has been shown in many previous experiments, see Colin F. Camerer (2003) for a book-length summary. This part of their conclusion is solid but is only original in its emphasis on the sampling and impulse balance models.

[4] Indeed, impulse balance equilibrium (with loss aversion) outperforms all other stationary concepts (without loss aversion). Once the other stationary concepts are augmented with loss aversion, they perform better than impulse balance equilibrium (see Figure 6 below).

reactions to foregone losses and gains is not innocuous. For the game designed by Goeree et al., it is the assumption of loss aversion that makes impulse balance equilibrium predict well.[5] As we show below, if the other concepts are augmented with loss aversion they predict behavior quite well (and even better than impulse balance equilibrium).

The second class of games that discriminate among concepts are asymmetric 2×2 matching pennies games (e.g. Jack Ochs, 1995). We report new analyses using the data of Richard D. McKelvey, Palfrey, and Roberto A. Weber (2000). In these games, loss aversion plays no role since security levels are 0 and payoffs are non-negative. We find that impulse balance equilibrium fits the same as QRE and somewhat worse than action-sampling and payoff-sampling. These two re-analyses of older data take up the search for games that discriminate better among stationary concepts that SC called for, and show that the loss-aversion built into impulse balance equilibrium accounts for some of that concept's success.

2. Re-Examining the SC Results

Table 1 shows data averages and model predictions for each of the 12 games. This table, and all subsequent tables and figures report corrections of their results in a visual form identical to their originals. The bold numbers indicate discrepancies between our results and those of SC. In particular, we find (i) a different impulse balance prediction for Game 1, (ii) a different data average for Game 3, (iii) a different optimal sample size ($n = 12$) and, hence, different predictions for action-sampling equilibrium (see Figure 1 for the mean-squared distances by sample size), and (iv) vastly different predictions for the QRE model: the precision parameter we estimate using the mean-squared distance objective function is $\lambda = 1.05$, much lower than the estimate reported by SC ($\lambda = 8.84$).[6]

At this lower value of λ, the QRE predictions (see Table 1) are much different from Nash predictions and much closer to the data. The improved fit is illustrated by Figure 2, which shows data averages and model predictions and parallels Figure 8 in SC. Using an 'ocular metric' suggests that the predictions of the alternative models are remarkably close to each other and to the data averages. To quantify this we also computed the sample variance and theory-specific variance as in SC, which are shown in Figure 3 (cf. Figure 12

[5]Following Ockenfels and Selten (2005), we estimated a one-parameter extension of impulse balance equilibrium where the weight for gains is fixed to be 1 but the weight for losses is a free parameter, γ. The estimations yield $\gamma = 2.07$ and the improvement in loglikelihood when γ is fixed at 2 is only 0.6%. In other words, the degree of loss aversion ($\gamma = 2$) that is hardwired into the impulse balance equilibrium concept is nearly optimal for the data set considered.

[6]Using maximum-likelihood techniques yields an estimate $\lambda = 0.99$.

Table 1 – Five Stationary Concepts Together with the Observed Relative Frequencies for Each of the Experimental Games.

		Nash	QRE (**λ=1.05**)	Action-sampling (**n=12**)	Payoff-sampling (n=6)	Impulse Balance	Observed average of 12 observations
Game 1	U	0.091	**0.042**	**0.090**	0.071	**0.068**	0.079
	L	0.909	**0.637**	**0.705**	0.643	0.580	0.690
Game 2	U	0.182	**0.154**	**0.193**	0.185	0.172	0.217
	L	0.727	**0.579**	**0.584**	0.569	0.491	0.527
Game 3	U	0.273	**0.168**	**0.208**	0.152	0.161	**0.163**
	L	0.909	**0.770**	**0.774**	0.771	0.765	0.793
Game 4	U	0.364	**0.275**	**0.302**	0.285	0.259	0.286
	L	0.818	**0.734**	**0.719**	0.726	0.710	0.736
Game 5	U	0.364	**0.307**	**0.329**	0.307	0.297	0.327
	L	0.727	**0.657**	**0.643**	0.654	0.628	0.664
Game 6	U	0.455	**0.417**	**0.426**	0.427	0.400	0.445
	L	0.636	**0.607**	**0.596**	0.597	0.600	0.596
		Nash	QRE (**λ=1.05**)	Action-sampling (**n=12**)	Payoff-sampling (n=6)	Impulse Balance	Observed average of 6 observations
Game 7	U	0.091	**0.042**	**0.090**	0.060	0.104	0.141
	L	0.909	**0.637**	**0.705**	0.691	0.634	0.564
Game 8	U	0.182	**0.154**	**0.193**	0.222	0.258	0.250
	L	0.727	**0.579**	**0.584**	0.602	0.561	0.587
Game 9	U	0.273	**0.168**	**0.208**	0.154	0.188	0.254
	L	0.909	**0.770**	**0.774**	0.767	0.764	0.827
Game 10	U	0.364	**0.275**	**0.302**	0.308	0.304	0.366
	L	0.818	**0.734**	**0.719**	0.730	0.724	0.700
Game 11	U	0.364	**0.307**	**0.329**	0.338	0.354	0.331
	L	0.727	**0.657**	**0.643**	0.650	0.646	0.652
Game 12	U	0.455	**0.417**	**0.426**	0.404	0.466	0.439
	L	0.636	**0.607**	**0.596**	0.599	0.604	0.604

Note: λ is logit precision parameter, n is the optimal sampling size for action or payoff sampling.

in SC).

SC evaluate the stationary concepts using data from both the first 100 periods and final 100 periods (as in their Figure 13). Our correction to their Figure 13 is Figure 4, which displays the theory-specific variances for the different concepts (excluding Nash) by the first and last blocks of 100 periods and for all 200 periods (correcting their Figure 12). It is notable that all the models fit substantially better in the last block than in the first block, as one would hope for reasonable concepts of stationary behavior (which are not necessarily designed to explain early behavior). It is also the case that impulse balance equilibrium is the best model in the first block of 100 periods, the worst in the second block of 100 periods, and is best using all periods.[7] It is an interesting question how model accuracy in

[7]This conclusion is different from what is concluded from SC's Figure 13, because of the corrections to both QRE and action-sampling, which improve their fit especially in the last block of 100 periods.

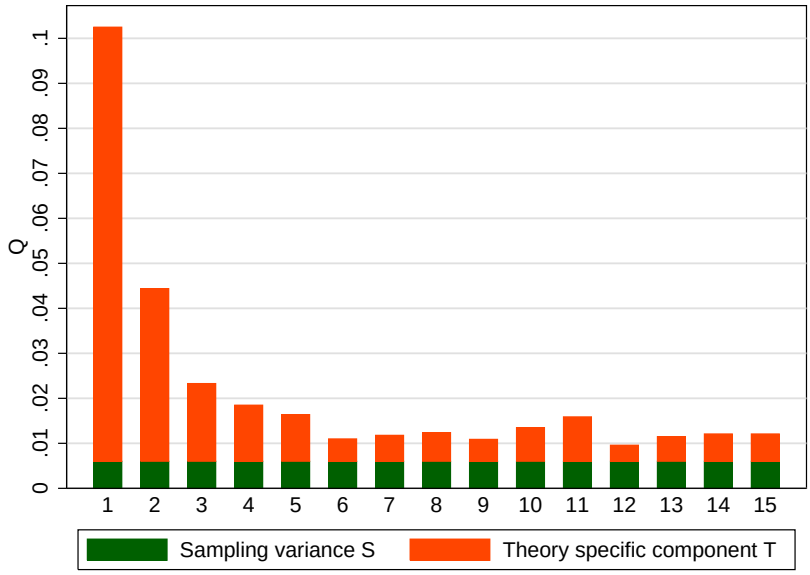

FIGURE 1. OVERALL MEAN SQUARED DISTANCES FOR THE ACTION-SAMPLING EQUILIBRIA WITH DIFFERENT SAMPLE SIZES (CF. SC FIGURE 9).

early, late, and all periods should be used to judge how well a stationary concept explains behavior.

To test for significant differences, SC report ten pairwise comparisons of the five different models based on the Wilcoxon matched-pairs signed-rank test. Each model generates a squared deviation (between observed and predicted frequencies) for each session, and the Wilcoxon test is applied to the differences in these squared deviations across models. Table 2 is an updated version correcting SC's pairwise model comparisons (see their Table 3). The model with the lowest rank-sum mean squared deviation is in the top row and the model with the highest rank-sum mean squared deviation is in the bottom row. The entries display rounded p-values for two-tailed Wilcoxon matched-pairs signed-rank tests for pairs of models, reported separately for constant-sum games, non-constant-sum games, and for all games (exactly as in their Table 3).[8] Combined, the various statistical tests confirm the general 'no difference' result suggested by Figure 2 – there is no clear ranking among the four stationary concepts that holds in all classes of games. However, all non-Nash models do much better than Nash and it is perhaps notable that action-sampling and payoff-sampling do better than QRE when all games are combined.

[8]The two entries below the diagonal in the impulse-balance row indicate that impulse balance equilibrium does significantly better than payoff sampling and logit-QRE for the non-constant sum games, but does worse overall (and significantly worse for the constant sum games).

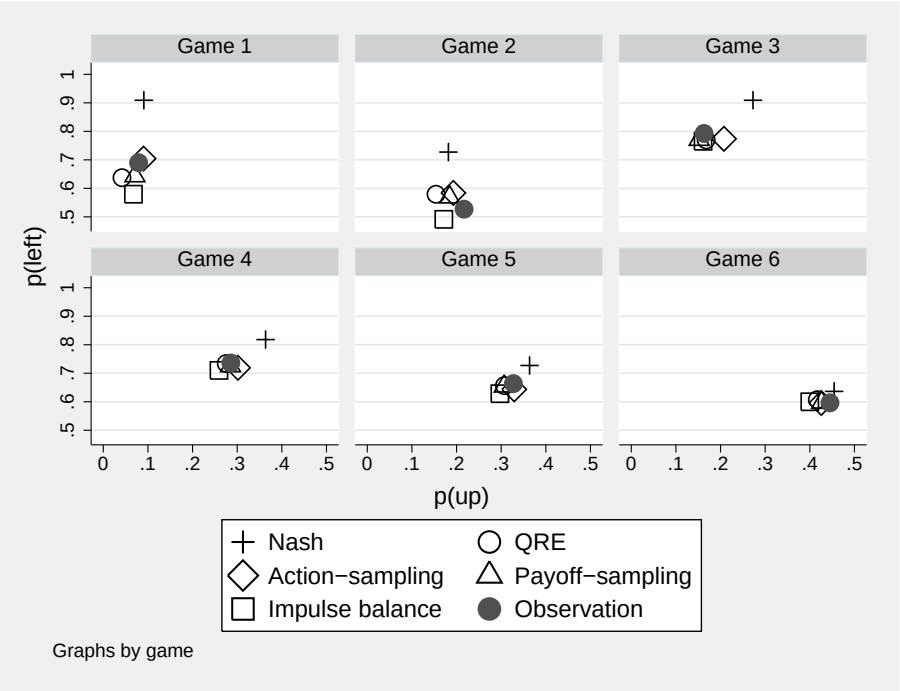

FIGURE 2. VISUALIZATION OF THE THEORETICAL EQUILIBRIA AND THE OBSERVED AVERAGE IN THE CONSTANT SUM GAMES (CF. SC FIGURE 8).

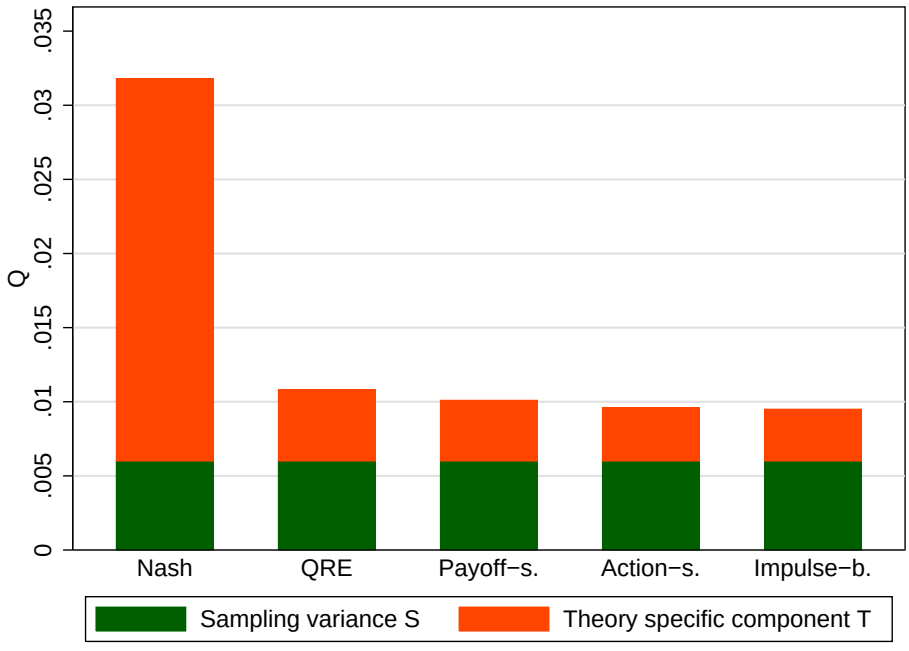

FIGURE 3. OVERALL MEAN SQUARED DISTANCES OF THE FIVE STATIONARY CONCEPTS COMPARED TO THE OBSERVED AVERAGE (CF. SC FIGURE 12).

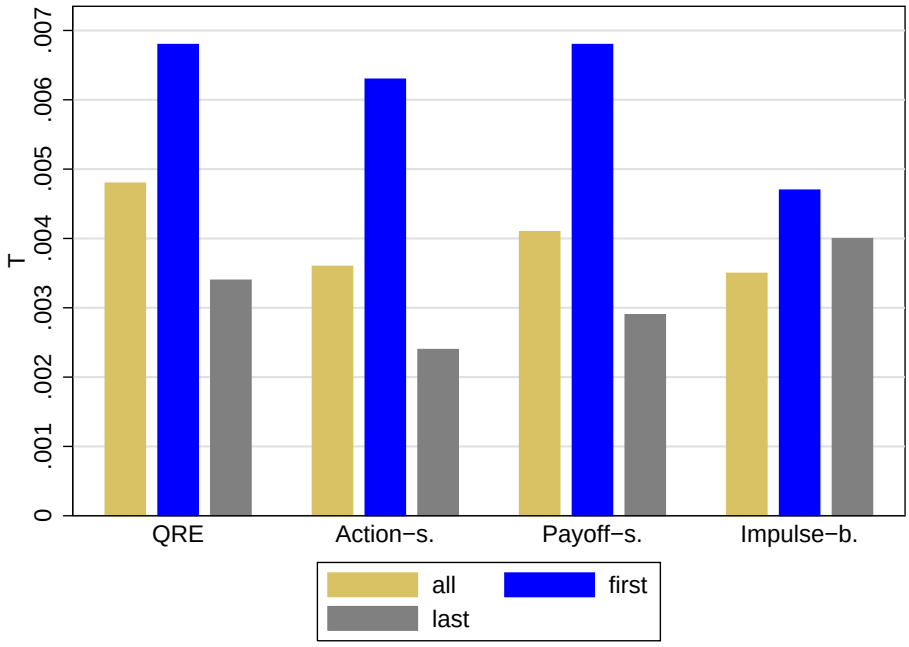

FIGURE 4. THEORY SPECIFIC SQUARED DISTANCES OF THE FIVE STATIONARY CONCEPTS COMPARED TO THE OBSERVED AVERAGE BY BLOCKS OF 100 PERIODS (CF. SC FIGURE 13).

As noted by a referee, the Wilcoxon test makes some distributional assumptions about the differences in squared deviations that are not necessarily true for the SC data. Therefore, we also performed two other non-parametric tests that relax these assumptions: the Kolmogorov-Smirnov two-sample test (KS) and a robust rank-order test (Fligner and Policello, 1982). When the KS test is applied to the ten different pairs of models, the null hypothesis of identical distributions cannot be rejected except when comparing action sampling and impulse balance equilibrium (at the 5% level) and when comparing any of the non-Nash models with Nash. The robust rank-order test also shows that none of the pairwise comparisons are significant except when comparing Nash to any of the non-Nash models.

To summarize, the SC design does not differentiate sharply among the stationary concepts they consider. The more robust non-parametric tests show this property most clearly (there are no general significant differences among non-Nash models). As noted in the Introduction, an extension to games which *do* differentiate well across concepts is therefore of interest in addressing the central point of the SC paper, which is comparison of stationary models.

Table 2 – P-values in Favor of Row Concepts, Two-Tailed Matched-Pairs Wilcoxon Signed Rank Test, n=108 (Rounded to the Next Higher Level Among 0.1 Percent, 0.2 Percent, 0.5 Percent, 1 Percent, 2 Percent, 5 Percent and 10 Percent).

	Action-sampling equilibrium	Payoff-sampling equilibrium	Quantal response equilibrium	Impulse balance equilibrium	Nash equilibrium
Action-sampling equilibrium		**n.s.** *n.s.* n.s.	**2 percent** *10 percent* 10 percent	**n.s.** *5 percent* n.s.	**0.1 percent** *0.1 percent* 0.1 percent
Payoff-sampling equilibrium			**5 percent** *0.1 percent* n.s.	**n.s.** *0.1 percent*	**0.1 percent** *0.1 percent* 0.1 percent
Quantal response equilibrium				**n.s.** *0.1 percent*	**0.1 percent** *0.1 percent* 1 percent
Impulse balance equilibrium					**0.1 percent** *0.1 percent* 0.1 percent
		5 percent	0.5 percent		

Above: All 108 Experiments; *Middle: 72 Constant-Sum Game Experiments*; Below: 36 Non-Constant Sum Game Experiments.

3. Differentiating Stationary Concepts in Other Data Sets

Goeree, Holt, and Palfrey (2003) designed the game in the left panel of Figure 5 to illustrate the limitations of QRE in terms of explaining behavior when other factors, such as risk aversion, are likely to play a role. In particular, both players have a "safe" choice that rewards either 160 or 200 and a "risky" choice that rewards either 10 or 370. Goeree et al. (2003) prove that in *any* quantal response equilibrium (logit or otherwise), the column player's probability of playing Right is greater than 0.5. Risk aversion, however, will naturally steer players towards the safer option of playing Left.

In the experiment, the aggregate choice frequencies were 65% for Left and 47% for Up, which contradicts risk-neutral QRE predictions. To compute the impulse-balance equilibrium of the game, note that the max-min payoff is 160 for both players. Subtracting 160 from all payoffs and multiplying by 2 if the resulting number is negative, yields the transformed game in the right panel of Figure 5. The condition that, for both players, the expected impulses even out yields: $300\,p_D\,q_R = 170\,p_U\,q_L$ and $300\,p_U\,q_R = 170\,p_D\,q_L$, which implies that $p_U = \frac{1}{2}$ and $q_L = \frac{30}{47} \approx 0.64$. Impulse balance equilibrium fits almost perfectly!

Keep in mind that in impulse balance equilibrium the response to perceived losses (rel-

	Left	Right
Up	200, 160	160, 10
Down	370, 200	10, 370

	Left	Right
Up	40, 0	0, -300
Down	210, 40	-300, 210

FIGURE 5. A MATCHING PENNIES GAME WITH "SAFE" (200/160) AND "RISKY" (370/10) CHOICES (LEFT) AND THE TRANSFORMED GAME (RIGHT).

ative to the max-min reference point) is twice as large as the response to gains. The authors are very clear that this asymmetry is a fixed feature of the theory, although in principle it could be treated as a free parameter (as, e.g., Ockenfels and Selten, 2005, did). Indeed, if losses and gains were weighed equally, the relevant conditions would be: $150\,p_D\,q_R = 170\,p_U\,q_L$ and $150\,p_U\,q_R = 170\,p_D\,q_L$, which implies that $p_U = \frac{1}{2}$ and $q_L = \frac{15}{32} \approx 0.47$. In other words, without the crucial loss aversion feature, the impulse balance equilibrium predictions are on the wrong side of 0.5 just like the risk-neutral QRE predictions reported by Goeree, Holt, and Palfrey (2003). The rightmost bars in Figure 6 show the theory-specific mean-squared deviations for impulse balance, with and without loss aversion. The other pairs of bars display the analogous results for Nash and non-Nash models – the latter do better than impulse balance equilibrium once they are also augmented with loss aversion (weighing losses twice as much as gains). Clearly, it is the loss aversion assumption that drives the goodness of fit for this game across all theories. It is true that only impulse balance equilibrium has loss aversion hardwired into it (Selten, Abbink, and Cox, 2005), but Figure 6 shows that adding loss aversion to the other theories (using the fixed value of two for the loss aversion parameter) improves fit dramatically.

3.1. Asymmetric Matching Pennies Games

A second test of the stationary concepts is provided by the experiments of McKelvey, Palfrey, and Weber (2000). They used games with an "asymmetric matching pennies" structure (Ochs, 1995), shown in Figure 7. The Row player earns a positive amount if the players match on "Heads" or "Tails" (and then the Column player earns nothing), or the Column player earns a positive amount if the players mismatch (and then the Row player earns nothing). McKelvey et al. (2000) consider four related games: in game A, $X = 9$; in game D, $X = 4$; game B payoffs are the same as game A's except Column payoffs are multiplied by 4; game C payoffs are the same as game A's except all payoffs are multiplied by 4.

To compute the impulse balance equilibria for these games note that the max-min payoff is 0 for both players (as is the second-lowest payoff) so the transformed games are equivalent

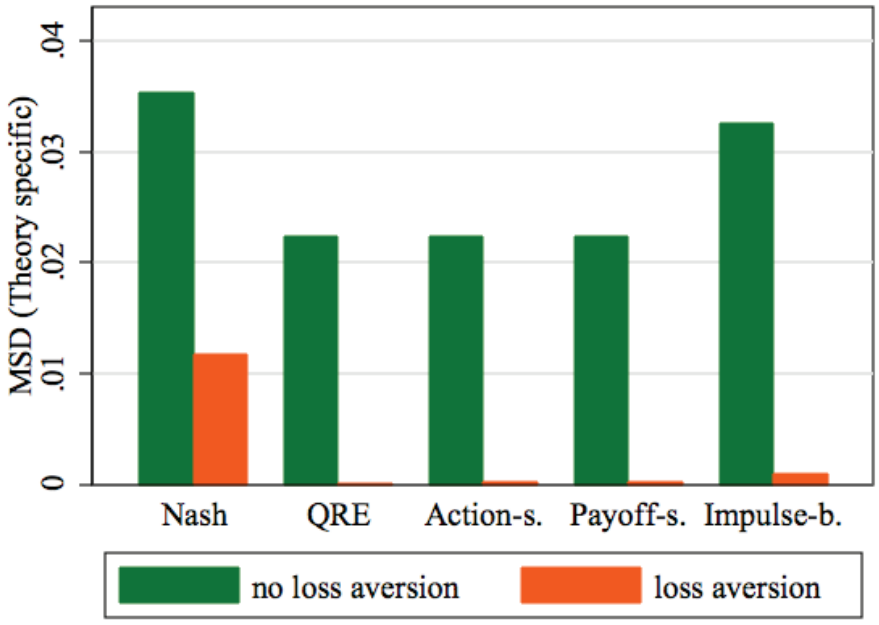

FIGURE 6. THEORY-SPECIFIC MEAN-SQUARED DISTANCES FOR GAME 4 FROM GOEREE ET AL. (2003) FOR MODELS WITH AND WITHOUT LOSS AVERSION.

to the original games. A simple calculation shows that for the game in Figure 6, the impulse balance equilibrium predictions for the Row and Column players are[9]

$$p_H = \frac{\sqrt{X}}{1+\sqrt{X}}, \qquad q_H = \frac{1}{1+\sqrt{X}}. \tag{1}$$

Since multiplicative factors drop out of the impulse balance equilibrium calculations, the predictions for games A, B, and C are identical: $p_H = 0.75$ for Row and $q_H = 0.25$ for Column, while for game D we have $p_H = 0.67$ for Row and $q_H = 0.33$ for Column.

The aggregate choice frequencies observed in the experiments are shown in Table 3 together with the predictions of the five stationary concepts (estimating any free parameters across all four games, with best-fitting parameters shown at the top of Table 3). The rightmost column reports the number of sessions of each game.[10] Wilcoxon signed-rank

[9]Interestingly, the predictions in (1) are identical to those obtained from a "Luce"-type quantal response equilibrium where choice probabilities are proportional to expected payoffs, e.g.

$$p_H = \frac{\pi_H^R}{\pi_H^R + \pi_T^R}, \qquad q_H = \frac{\pi_H^C}{\pi_H^C + \pi_T^C}.$$

The expected payoffs on the right side depend on the choice probabilities: $\pi_H^R = q_H X$, $\pi_T^R = 1 - q_H$ and $\pi_H^C = 1 - p_H$, $\pi_T^C = p_H$. It is straightforward to show the fixed-point probabilities are those in (1).

[10]In the McKelvey, Palfrey, and Weber (2000) experiments, subjects played 50 periods using one of their

	Heads	Tails
Heads	X, 0	0, 1
Tails	0, 1	1, 0

FIGURE 7. AN ASYMMETRIC MATCHING PENNIES GAME.

tests across McKelvey, Palfrey, and Weber's sessions indicate that all non-Nash models are significantly more accurate than Nash, the action-sampling and payoff-sampling models are more accurate than QRE and impulse balance equilibrium, and the latter two are equally accurate.[11]

4. Conclusion

This comment corrects and re-examines some of the results reported by SC. Correcting for errors in estimating two of the five stationary concepts they consider, QRE and action-sampling equilibrium, it appears that their design does not differentiate among the non-Nash stationary concepts that were considered. They also suggest it is desirable to create games which discriminate among these non-Nash theories, a direction which we pursue by reporting two new analyses. We first tested these concepts further by using data from previous laboratory experiments on a game constructed to show that QRE can predict poorly. Applying all five stationary concepts to those data, with and without loss-aversion, shows that the loss-aversion that is a fixed feature of impulse balance equilibrium is crucial for its explanatory power in this particular game. This result extends our understanding of which modeling features of various theories are responsible for accurate fit. We also tested the stationary concepts on four variants of matching pennies games. In these games, all theories fit much better than Nash but action-sampling and payoff-sampling fit a little better than the other non-Nash theories.

The results reported in SC and here on how closely many of these theories approximate behavior also suggests the possibility that the different approaches might share some similar hidden features. For some games, impulse balance equilibrium *coincides* with a specific quantal response equilibrium model (footnote 9). Furthermore, by varying the sample size in the action sampling equilibrium from one to infinity, the implied behavior ranges from

game forms and then played another 50 periods using another one of their game forms. In the analysis reported here, we consider only the first 50 periods of play.

[11]The models are ranked: payoff-sampling $\sim_{p=0.55}$ action-sampling $\succ_{p=0.02}$ impulse balance $\sim_{p=0.11}$ QRE $\succ_{p=0.01}$ Nash. These comparisons are confirmed by the Kolmogorov-Smirnov and Fligner-Policello tests.

Table 3 – Five Stationary Concepts Together with the Observed Relative Frequencies for Each of the Experimental Games in McKelvey, Palfrey and Weber (2000).

	Nash	QRE (λ=3.62)	Action-sampling (n=3)	Payoff-sampling (n=3)	Impulse Balance	Observed average	Number of observations
U	0.500	0.760	0.643	0.625	0.750	0.648	3
L	0.100	0.132	0.291	0.276	0.250	0.245	
U	0.500	0.573	0.643	0.625	0.750	0.627	2
L	0.100	0.108	0.291	0.276	0.250	0.300	
U	0.500	0.575	0.643	0.625	0.750	0.608	2
L	0.100	0.102	0.291	0.276	0.250	0.218	
U	0.500	0.661	0.643	0.625	0.667	0.643	1
L	0.200	0.237	0.291	0.276	0.333	0.343	
MSD	0.0441	0.0256	0.0057	0.0054	0.0153		

purely random to Nash behavior in the 2×2 games studied here. This is akin to varying the precision parameter in a QRE model, and as such, the two models trace out different one-dimensional curves in the two-dimensional unit square corresponding to players' choice probabilities. (And payoff sampling yields yet another such curve.) Evaluating theories via a simple "horse race" is simply asking which of these curves comes closest to the observed data points. By allowing oneself some degree of freedom in the choice of underlying modeling assumptions (e.g. the logit specification of QRE) or behavioral assumptions (e.g. the loss aversion hardwired into impulse balance equilibrium), such curve-fitting comparisons become difficult to interpret.

As a result, the value of these models cannot simply be measured by their ability to come out ahead in a statistical comparison of fit. It is helpful if models are analytically tractable, fit a wide range of games with comparable parameter values using different fit and prediction measures, and perhaps are even consistent with non-choice measures (such as eyetracking). Therefore, it is necessary to continue to search for combinations of parsimonious theoretical features which predict well across many games and measures.

References

Camerer, Colin F. (2003), *Behavioral Game Theory: Experiments on Strategic Interaction*, Princeton: Princeton University Press.

Capra, C. Monica, Rosario Gomez, Jacob K. Goeree, and Charles H. Holt (2002), "Learning and Noisy Equilibrium Behavior in an Experimental Study of Imperfect Price Competition," *International Economic Review*, **43**(3), 613-636.

Fligner, M. A. and Policello, G. E. (1981) "Robust Rank Procedures for the Behrens-Fisher Problem," *Journal of the American Statistical Association*, **76**(373), 162-168.

Goeree, Jacob K. and Charles A. Holt (2002) "Learning in Economics Experiments," *Encyclopedia of Cognitive Science*, Volume **2**, L. Nagel, ed., London: Nature Publishing Group, McMillan, 1060-1069.

Goeree, Jacob K., Charles A. Holt, and Thomas R. Palfrey (2003) "Risk Averse Behavior in Generalized Matching Pennies Games," *Games and Economic Behavior*, **45**(1), 97-113.

McKelvey, Richard D., Thomas R. Palfrey, and Roberto A. Weber (2000) "The Effects of Payoff Magnitude and Heterogeneity on Behavior in 2×2 Games with a Unique Mixed Strategy Equilibrium," *Journal of Economic Behavior and Organization*, **42**(4), 523-548.

Ockenfels, Axel and Reinhard Selten (2005) "Impulse Balance Equilibrium and Feedback in First Price Auctions," *Games and Economic Behavior*, **51**(1), 155-170.

Ochs, Jack (1995) "Games with a Unique Mixed-Strategy Equilibrium: an Experimental Study," *Games and Economic Behavior*, **10**, 202-217.

Selten, Reinhard, Klaus Abbink, and Ricarda Cox (2005) "Learning Direction Theory and the Winners Curse," *Experimental Economics*, **8**(1), 5-20.

Selten, Reinhard and Thorsten Chmura (2008) "Stationary Concepts for Experimental 2×2 Games," *American Economic Review*, **98**(3), 938-966.