

Stationary Concepts for Experimental 2×2 Games: Comment

By CHRISTOPH BRUNNER, COLIN F. CAMERER, AND JACOB K. GOEREE*

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 two different restrictions: that players 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 do not 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 forgone payoffs. If the option not chosen would have yielded higher payoffs, then there is an "impulse" to change (and, importantly, "losses" of forgone 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 (2008, p. 962) conclude that Nash and QRE fit worse than the other three concepts. They write:

*Brunner: Alfred-Weber-Institute, University of Heidelberg, Bergheimerstrasse 58, 69115 Heidelberg, Germany (e-mail: cbrunner@uni-hd.de); Camerer: Division of the Humanities and Social Sciences, California Institute of Technology, Mail code 228-77, Pasadena, CA 91125 (e-mail: camerer@hss.caltech.edu); Goeree: Institute for Empirical Research in Economics, University of Zürich, Blümlisalpstrasse 10, CH-8006, Zürich, Switzerland (e-mail: jacob.goeree@gmail.com). 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 European Research Council (ERC Advanced grant, ESEI-249433).

¹ SC (2008, p. 939) write "The concept has been developed by one of the authors (Selten). As far as we know, it cannot be found in the literature." However, Goeree and Charles A. Holt (2002) introduced the concept of "stochastic learning equilibrium," which describes the long-run outcome when players better respond to a weighted sample of their opponent's previous actions. In other words, the models are similar because one uses a finite sample (bounded memory) and another uses a discounted sample (fading memory), though neither model is nested within the other, and stochastic learning equilibrium allows for both best and better responses. The stochastic learning equilibrium is also clearly a stationary concept, *not* a model of learning, intended to describe long-run outcomes.

"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."

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 weakens the most novel part of their original conclusion, i.e., "all the relevant comparisons are highly significant," and implies a weaker conclusion: the sampling theories do better in some (but not all) comparisons, and QRE does not fit worse (or better) than impulse balance equilibrium.³

Fit measures and statistical tests show that the four non-Nash models are about equally accurate. SC (2008, p. 965) 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* (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 reactions to forgone losses and gains is not innocuous. For the game designed by Goeree, Holt, and Palfrey (2003), it is the assumption of loss aversion that makes impulse balance equilibrium predict well.⁵

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 hundreds of previous experiments; see 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).

⁵ Following Axel Ockenfels and Reinhard 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 log-likelihood when γ is fixed at 2 is only 0.6 percent. In other

TABLE 1—FIVE STATIONARY CONCEPTS TOGETHER WITH THE OBSERVED RELATIVE FREQUENCIES FOR EACH OF THE EXPERIMENTAL GAMES

| | | Nash | QRE ($\lambda = 1.05$) | 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 ($\lambda = 1.05$) | 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: γ = is the logit precision parameter, n the optimal sampling size for action or payoff sampling.

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 zero and payoffs are positive. 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.

I. Reexamining 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

words, the degree of loss aversion ($\gamma = 2$) that is hardwired into the impulse balance equilibrium concept is nearly optimal for the dataset considered.

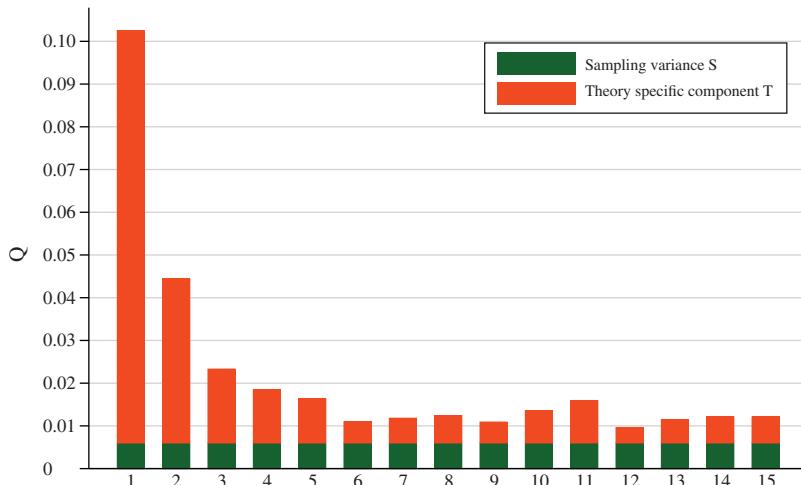


FIGURE 1. OVERALL MEAN SQUARED DISTANCES FOR THE ACTION-SAMPLING EQUILIBRIA WITH DIFFERENT SAMPLE SIZES (cf. SC 2008, Figure 9)

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 (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 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 in SC).

SC (2008) evaluate the stationary concepts using data from the first 100 periods and final 100 periods (as in their Figure 13). Our correction to their Figure 13 is Figure 4, which displays theory-specific variances for the different concepts (excluding Nash) by the first and last blocks of 100 periods and for all 200 periods

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

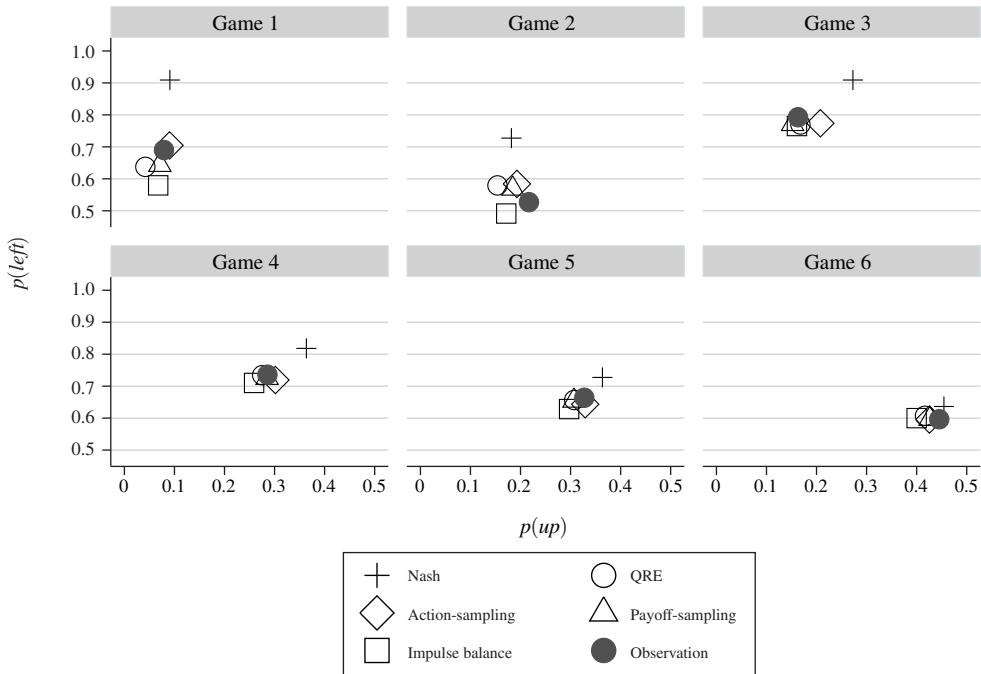


FIGURE 2. VISUALIZATION OF THEORETICAL EQUILIBRIA AND THE OBSERVED AVERAGE IN THE CONSTANT SUM GAMES (cf. SC 2008, Figure 8)

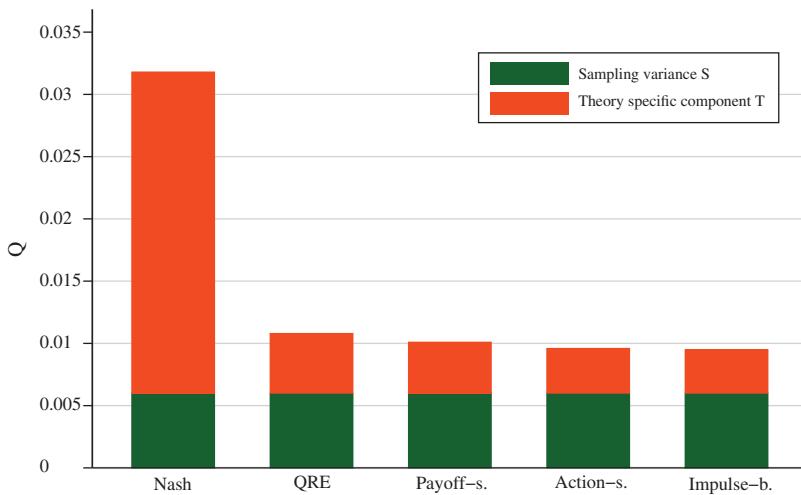


FIGURE 3. OVERALL MEAN SQUARED DISTANCES OF THE FIVE STATIONARY CONCEPTS COMPARED TO THE OBSERVED AVERAGE (cf. SC 2008, Figure 12)

(correcting their Figure 12). It is notable that all 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

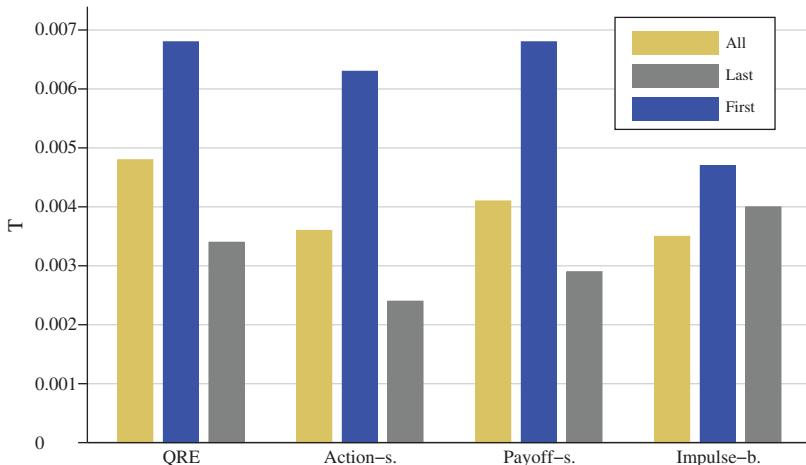


FIGURE 4. THEORY SPECIFIC SQUARED DISTANCES OF THE FIVE STATIONARY CONCEPTS COMPARED TO THE OBSERVED AVERAGE BY BLOCKS OF 100 PERIODS (*cf. SC Figure 13*)

all periods.⁷ It is an interesting question how model accuracy in 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 matched-pairs signed-rank test. Each model generates a squared deviation (between observed and predicted frequencies) for each of the 108 sessions, 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 (compare with their Table 3). The top-to-bottom order of the models is the same as in their original table. 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 “no difference” result suggested by Figure 2—there is no clear ranking among the four non-Nash models that holds across all games. The no difference result is all the more remarkable as the tests are based on 108 matched-pair observations. As expected the non-Nash models all 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.

In their reply to this comment, Selten, Chmura, and Sebastian J. Goerg (2011)—henceforth, SCG—now suggest implicitly that the Wilcoxon test they used earlier is problematic because of the assumption of symmetry around the median (although the assumption seems to be satisfied empirically⁹). SCG (2011) now propose using the Fisher-Pitman test, which is newly reported in their reply. The only difference

⁷This conclusion is different from what is concluded from SC (2008)'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.

⁸Note that impulse-balance equilibrium does significantly better than payoff sampling and logit-QRE for the non-constant sum games, but it does significantly worse for the constant-sum games and worse overall.

⁹Ronald H. Randles et al. (1980) develop a test for whether data are symmetrically distributed around an unknown median. We applied their test to the matched pairs of squared deviations across all possible theory pairs, and none of the resulting *p*-values were at all close to significant (ranging from 0.32 to 0.89). So the assumption of symmetry is reasonable and adds an additional empirical justification to SC's initial use of the Wilcoxon test.

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)

| | Impulse balance equilibrium | Payoff-sampling equilibrium | Action-sampling equilibrium | Quantal response equilibrium | Nash equilibrium |
|------------------------------|--|-----------------------------|-----------------------------|---|--|
| Impulse balance equilibrium | | | | | 0.1 percent <i>0.1 percent</i> |
| Payoff-sampling equilibrium | n.s. <i>0.1 percent</i> | 5 percent | | 0.5 percent 5 percent <i>0.1 percent</i> | 0.1 percent 0.1 percent <i>0.1 percent</i> |
| Action-sampling equilibrium | n.s. <i>5 percent</i> <i>n.s.</i> | n.s. <i>n.s.</i> | | 2 percent <i>10 percent</i> n.s. 10 percent | 0.1 percent <i>0.1 percent</i> <i>0.1 percent</i> |
| Quantal response equilibrium | n.s. <i>0.1 percent</i> | | | | 0.1 percent <i>0.1 percent</i> 1 percent |

Notes: Above: all 108 experiments are in bold; middle: 72 constant-sum game experiments; below: 36 non-constant sum game experiments

in the corrected Wilcoxon (W) results and the newly reported Fisher-Pitman (F-P) results across all games is that action-sampling equilibrium is now only weakly more accurate than QRE at $p < 0.1$ (F-P) rather than at $p < 0.02$ (W). In subsets of games, there are some minor differences in the p -values at which differences are significant (p -values generally are higher so results are weaker using the F-P test). QRE is significantly more accurate than impulse balance in the constant-sum game subset (at $p < 0.001$ using Wilcoxon), but using the F-P test, this result is erased, and impulse balance is also then more accurate than QRE in the non-constant sum game subset (at $p < 0.01$).

To summarize, neither the Wilcoxon test originally applied by SC nor the weaker Fisher-Pitman test differentiate very sharply among the stationary concepts. 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 the comparison of stationary models.

II. Differentiating Stationary Concepts in Other Datasets

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, Holt, and Palfrey (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 percent for Left and 47 percent 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

| | Left | Right | | Left | Right |
|------|----------|---------|------|---------|-----------|
| Up | 200, 160 | 160, 10 | Up | 40, 0 | 0, -300 |
| Down | 370, 200 | 10, 370 | 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*).

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: $150p_Dq_R = 85p_Uq_L$ and $150p_Uq_R = 85p_Dq_L$, which implies that $p_U = 1/2$ and $q_L = 30/47 \approx 0.64$. Impulse balance equilibrium fits almost perfectly!

Keep in mind that in impulse balance equilibrium the response to perceived losses (relative 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: $150p_Dq_R = 170p_Uq_L$, and $150p_Uq_R = 170p_Dq_L$, which implies that $p_U = 1/2$ and $q_L = 15/32 \approx 0.47$. 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).¹⁰ 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 (Reinhard Selten, Klaus Abbink, and Ricarda 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.

A. Asymmetric Matching Pennies Games

A second test of the stationary concepts is provided by the experiments of McKelvey, Palfrey, and Weber (2000) based on games with an “asymmetric matching pennies” structure (Ochs 1995); see 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 the Row player earns nothing). McKelvey, Palfrey, and Weber (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.

¹⁰The model estimates without loss aversion are: $n = 1$ for action sampling, $n = 1$ for payoff sampling, and $\lambda = 0$ for logit-QRE. The logit-QRE estimate shows that the model cannot by itself explain behavior in the game of Figure 5, which was the main motivation for its design. With loss aversion the estimates are: $n = 4$ for action sampling, $n = 4$ for payoff sampling, and $\lambda = 0.024$ for logit-QRE.

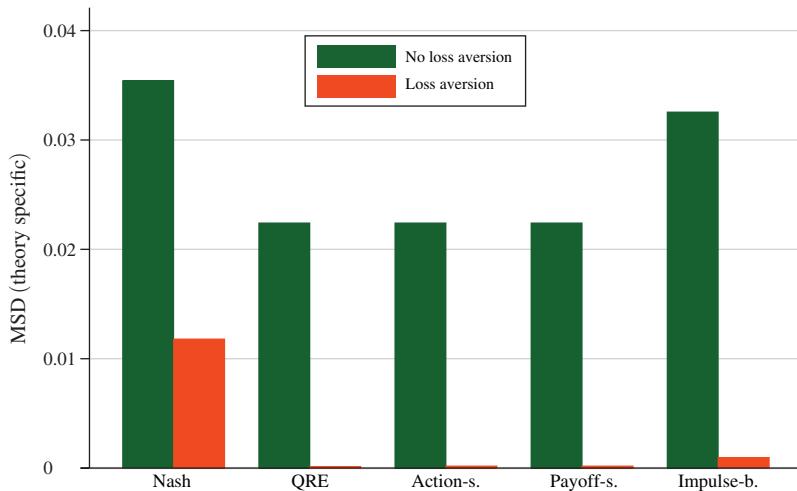


FIGURE 6. THEORY-SPECIFIC MEAN-SQUARED DISTANCES FOR GAME 4 FROM GOEREE, HOLT, AND PALFREY (2003) FOR MODELS WITH AND WITHOUT LOSS AVERSION

To compute the impulse balance equilibria for these games note that the max-min payoff is 0 for both players (as it is the second-lowest payoff), so the transformed games are equivalent to the original games. In other words, *loss aversion plays no role in these games*, which makes them ideal to further test the different stationary concepts.

A simple calculation shows that for the game in Figure 7, the impulse balance equilibrium predictions for the Row and Column players are¹¹

$$(1) \quad p_H = \frac{\sqrt{X}}{1 + \sqrt{X}}, \quad q_H = \frac{1}{1 + \sqrt{X}}.$$

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).¹² The rightmost column reports the number of sessions of each game.¹³ Using a Wilcoxon signed-rank test to evaluate differences in MSD across McKelvey, Palfrey, and Weber's sessions (number of observations shown in the rightmost column of Table 3) shows that all non-Nash models are significantly more accurate

¹¹ The impulse balance equations are $X(1 - p)q = p(1 - q)$ for the Row player and $pq = (1 - p)(1 - q)$ for the Column player.

¹² Recall that the games considered by McKelvey, Palfrey, and Weber (2000) are such that loss aversion plays no role, so all models are estimated without loss aversion.

¹³ In the McKelvey, Palfrey, and Weber (2000) experiments, subjects played 50 periods using one of their 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.

TABLE 3—FIVE STATIONARY CONCEPTS TOGETHER WITH THE OBSERVED RELATIVE FREQUENCIES
FOR EACH OF THE EXPERIMENTAL GAMES IN MCKELVEY, PALFREY, AND WEBER (2000)
WHERE LOSS AVERSION PLAYS NO ROLE (*all models are estimated without loss aversion*)

| Game | | Nash | QRE ($\lambda = 3.62$) | Action-sampling ($n = 3$) | Payoff-sampling ($n = 3$) | Impulse balance | Observed average | Number of observations |
|------|---|--------|-----------------------------|--------------------------------|--------------------------------|-----------------|------------------|------------------------|
| A | 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 | |
| B | 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 | |
| C | 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 | |
| D | 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 | | |

than Nash (at the 1 percent level), the action-sampling and payoff-sampling models are more accurate than QRE and impulse balance equilibrium (at the 2 percent level), and impulse balance equilibrium appears to have a nonnegligible advantage over QRE, but it does not reach conventional levels of significance in this dataset.

III. Conclusion

This comment corrects and reexamines 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.

One distinguishing element of impulse balance equilibrium vis-à-vis the other non-Nash models is that it is “parameter free,” since the loss-aversion coefficient is calibrated to 2. This can be a desirable feature from a theoretical viewpoint but makes the model less suitable for empirical applications. Ockenfels and Selten (2005), for example, introduce *weighted* impulse balance equilibrium to allow for a general ratio, γ , that measures the importance of upward and downward impulses in first-price auctions. They estimate that $\gamma = 3$, i.e., upward impulses that occur when the auction is lost are roughly three times larger than downward impulses that occur when the winner of the auction “left money on the table.” As Ockenfels and Selten (2005, p. 166) argue, “...there is no reason to assume that γ cannot change with

| | Heads | Tails |
|-------|-------|-------|
| Heads | X, 0 | 0, 1 |
| Tails | 0, 1 | 1, 0 |

FIGURE 7. AN ASYMMETRIC MATCHING PENNIES GAME

experiment design parameters..." Indeed, a survey of studies that estimate loss aversion coefficients from lab or field data shows that γ varies (from 1.5 to 4.5) across games, contexts, and subject pools; see Colin F. Camerer (2009).

Allowing for a one-parameter extension of the basic impulse balance model to improve empirical applicability is closely related to the introduction of the precision parameter in QRE and to sample size parameters in action- and payoff-sampling theories. Comparing these models with Nash equilibrium, and impulse balance with a weight of two, raises interesting questions about the advantages and disadvantages of parametric approaches. Certainly, both parametric and parameter-free approaches should be explored, but our view of parametric models of behavior in games is more optimistic than the view articulated by SCG in their reply to this comment. In SCG (2011, p. 1044) they state that "nonparametric concepts like the IBE [impulse balance equilibrium] have the advantage to serve as the basis of theoretical investigations just like NE [Nash equilibrium]." Of course, parametric models can also be investigated theoretically, and have been, many times. Indeed, many comparative static predictions from a wide range of models in all areas of economics are of this sort; they report how behavior should respond to parameter changes (e.g., how behavior varies with a risk-aversion or time-preference parameter). To be sure, those theoretical results will typically depend on specific parameter values, but they are theoretical results nonetheless. SCG (2011, p. 1043) also assert that "Moreover, it is obviously not possible to transfer parameter estimates for a small number of very similar games to wider classes of games." It is surely too pessimistic to declare such predictions "not possible." Making accurate new predictions is quite possible if parameter variation is not very wide, or if theory can eventually be developed to explain why parameters vary across games. In fact, if parameter variation reflects some fundamental aspect of players' cognition, such as experience, analytical skill, attentiveness, or memory, then parameters *should* vary across games. Then the only question is how reported estimates can be used to create good theory about how parameters vary, turning that variation from an inevitable bug into a valuable feature.

REFERENCES

- Camerer, Colin F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, NJ: Princeton University Press.
- Camerer, Colin F. 2009. "The Promise of Lab-Field Generalizability in Experimental Economics." Unpublished.
- Goeree, Jacob K., and Charles A. Holt. 2002. "Learning in Economics Experiments." In *Encyclopedia of Cognitive Science*. Vol. 2, ed. Lynn Nadel, 1060–69. London: Nature Publishing Group, McMillan.

- 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 Unique Mixed Strategy Equilibria." *Journal of Economic Behavior and Organization*, 42(4): 523–48.
- Ochs, Jack.** 1995. "Games with Unique, Mixed Strategy Equilibria: An Experimental Study." *Games and Economic Behavior*, 10(1): 202–17.
- Ockenfels, Axel, and Reinhard Selten.** 2005. "Impulse Balance Equilibrium and Feedback in First Price Auctions." *Games and Economic Behavior*, 51(1): 155–70.
- Randles, Ronald H., Michael A. Fligner, George E. Policello, and Douglas A. Wolfe.** 1980. "An Asymptotically Distribution-Free Test for Symmetry Versus Asymmetry." *Journal of the American Statistical Association*, 75(369): 168–72.
- Selten, Reinhard, Klaus Abbink, and Ricarda Cox.** 2005. "Learning Direction Theory and the Winner's 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): 938–66.
- Selten, Reinhard, Thorsten Chmura, and Sebastian J. Goerg.** 2011. "Stationary Concepts for Experimental 2×2 -Games: Reply." *American Economic Review*, 101(2): 1041–44.

This article has been cited by:

1. Reinhard Selten, , Thorsten Chmura, , Sebastian J. Goerg. 2011. Stationary Concepts for Experimental 2×2 Games: ReplyStationary Concepts for Experimental 2×2 Games: Reply. *American Economic Review* 101:2, 1041-1044. [[Abstract](#)] [[View PDF article](#)] [[PDF with links](#)]