American Economic Review 101 (April 2011): 1041–1044
http://www.aeaweb.org/articles.php?doi=10.1257/aer.101.2.1041

Stationary Concepts for Experimental 2×2 Games: Reply

By Reinhard Selten, Thorsten Chmura, and Sebastian J. Goerg*

The comment of Christoph Brunner, Colin F. Camerer, and Jacob K. Goeree (2011) identifies and corrects some computational errors in Selten and Chmura (2008). The first two authors of this reply deeply regret their computational mistakes and are grateful for the corrections. Nevertheless, we would like to add some short remarks on the comment.

I. Interpretation of the Corrected Results

In view of the objections concerning the distributional assumptions of the Wilcoxon test, we present the results of an alternative test, which does not depend on this assumption, namely the Fisher-Pitman permutation test for paired replicates (FP).[1] Table 1 gives the test results.

The concepts IBE, ASE, PSE, and QRE perform significantly better than NE, over all games and for each of the two types of games. Besides this, the FP test implies that impulse balance equilibrium (IBE), action-sampling equilibrium (ASE), and payoff-sampling equilibrium (PSE) perform about equally well. IBE is significantly more successful than ASE and PSE for nonconstant-sum games and the opposite is true for constant-sum games. Overall no significant difference between ASE, PSE, and IBE is observed. The performances of ASE and PSE dot not differ significantly for the constant- and nonconstant-sum games either.

No comparison of IBE, ASE, and PSE with quantal response equilibrium (QRE) is in favor of the latter one. In contrast to this, there are comparisons with QRE that are in favor of IBE, ASE, and PSE. IBE performs significantly better in the nonconstant-sum games, ASE performs significantly better over all games, and PSE performs significantly better over all games and for the constant-sum games.

Admittedly, this test (as well as the Wilcoxon sign-rank test) does not differentiate strongly among the four non-Nash concepts, but the assertion of no differentiation seems to be exaggerated. We cautiously summarize the results (which also hold for the Wilcoxon sign-rank test as reported by Brunner, Camerer, and Goeree 2011) as follows: the comparison among the parametric stationary concepts is clearly in favor of ASE and PSE and in disfavor of QRE. The nonparametric concept IBE performs

* Selten: BonnEconLab, Adenauerallee 24-42, 53113 Bonn, Germany (e-mail: rselten@uni-bonn.de); Chmura: University of Munich, LMU, Department of Economics, Seminar for Economic Theory, Ludwigstrasse 28, 0539 Muenchen, Germany, and BonnEconLab (e-mail: chmura@uni-bonn.de); Goerg (corresponding author): Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, 53113 Bonn, Germany, and BonnEconLab (e-mail: goerg@coll.mpg.de). Financial support by the Deutsche Forschungsgemeinschaft (DFG) is gratefully acknowledged.

[1] Refer to the online Appendix for a short explanation of the concept behind this test.

TABLE 1—TWO-SIDED SIGNIFICANCES IN FAVOR OF ROW CONCEPTS, MONTE-CARLO APPROXIMATION OF THE FISHER-PITMAN PERMUTATION TEST FOR PAIRED REPLICATES

	Impulse-balance equilibrium	Payoff-sampling equilibrium	Action-sampling equilibrium	Quantal response equilibrium	Nash equilibrium
Impulse balance equilibrium		**n.s.** — 1 percent	**n.s.** — 10 percent	**n.s.** *n.s.* 0.1 percent	**0.1 percent** *0.1 percent* 0.1 percent
Payoff-sampling equilibrium	**n.s.** *1 percent* —			**5 percent** *0.1 percent* n.s.	**0.1 percent** *0.1 percent* 0.2 percent
Action-sampling equilibrium	**n.s.** *10 percent* —	**n.s.** *n.s.* n.s.		**10 percent** *n.s.* n.s.	**0.1 percent** *0.1 percent* 0.1 percent
Quantal-response equilibrium					**0.1 percent** *0.1 percent* 1 percent

Notes: Rounded to the next higher level among 0.1 percent, 0.2 percent, 1 percent, 2 percent, 5 percent, and 10 percent. n.s.: not significant. Above: all 108 experiments; middle: 72 constant-sum game experiments; below: 36 nonconstant-sum game experiments.

significantly better than its nonparametric competitor (NE) and scores at least as well as the parametric concept QRE.

II. Differentiating Stationary Concepts in Other Data Sets

In their comment, Brunner, Camerer, and Goeree (2011) extend the comparison of the five stationary concepts to data from previous experimental studies. Important features of Selten and Chmura (2008) were the systematical variation of games, a large number of repeated rounds and the collection of a conclusive number of observations.[2] The data reevaluated by Brunner, Camerer, and Goeree (2011) meet none of these criteria.[3] Nevertheless, the extended comparisons do not yield results contradicting the corrected ones of Selten and Chmura (2008). Our summary for the corrected results of Selten and Chmura (2008) also holds for the newly investigated datasets: there are parametric concepts (ASE, PSE) that perform better than QRE, and there is one nonparametric concept (IBE) that performs at least as well as QRE.

A. *Reevaluation with a Matching Pennies Game*

Brunner, Camerer, and Goeree (2011) test the performance of the five concepts for game 4 of Goeree et al. (2003). As Figure 6 in section 3 of Brunner, Camerer, and Goeree (2011) shows, IBE performs much better for these data than the other four concepts. However, if loss-aversion, as incorporated in IBE, is added to the other concepts, then QRE, PSE, and ASE perform slightly better than IBE. Does this justify the conclusion that the success of IBE is solely driven by the incorporation of loss aversion? If incorporation of loss aversion improved NE, QRE, ASE, and PSE

[2] Refer to Figure 7 in Selten and Chmura (2008) to see the broad set of covered Nash equilibria. All games were repeated over 200 rounds. Overall 108 independent observations were gathered with a minimum of six per game.

[3] The studies by Goeree, Charles A. Holt, and Thomas R. Palfrey (2003) and McKelvey, Palfrey, and Roberto A. Weber (2000) did not investigate the issue of an objective comparison of stationary concepts; therefore these objections are not directed against the original studies.

in general, this should also be true for the Selten and Chmura (2008) data. In Selten and Chmura (2008), this question was investigated by Figure 11. The performances of NE, PSE, and ASE, contrary to the one of IBE, decrease significantly if loss aversion is incorporated, and this also holds true for the corrected data. If the same comparison is applied to QRE, no significant difference over all games is observed. In the constant-sum games QRE performs significantly better without incorporating loss aversion, while in the nonconstant-sum games it performs significantly better with loss aversion.[4] This suggests that the results of Brunner, Camerer, and Goeree (2011) cannot be generalized beyond game 4 of Goeree, Holt, and Palfrey (2003), and therefore they have no general implications.

In view of the results of the corrected Figure 11, additional properties other than loss aversion must contribute to the good fit of IBE. In this context, it is important to note that IBE does not involve any optimization, whereas all the other concepts have some aspect of optimization: NE optimizes against the other players' strategy, ASE against a sample of the other players' actions, PSE against sampled payoffs for own strategies, and QRE applies an error disturbed optimization against the same kind of behavior by the other players. In contrast to this, the balancing of expected impulses involves no element of optimization.

B. *Reevaluation with Asymmetric Matching Pennies Games*

In section 3.1 of Brunner, Camerer, and Goeree (2011), the stationary concepts are applied to four asymmetric matching pennies games (A, B, C, D) of McKelvey, Palfrey, and Weber (2000). Interestingly, the estimates of the QRE parameter and the sample sizes for ASE and PSE are very different from those for the SC-data. This raises the question whether it is reasonable to fit these parameters to such a small database (eight independent matching groups versus 108 independent matching groups in Selten and Chmura 2008). Moreover, it is obviously not possible to transfer parameter estimates for a small number of very similar games to wider classes of games.

III. The Aim of Comparing Stationary Concepts

Nash equilibrium and its refinements are very successful tools of economic analysis. For every game model with a unique Nash equilibrium, one gets a very clear prediction. Since, unfortunately, mixed Nash equilibrium is not a good predictor, it is necessary to develop an analytically tractable behavioral concept with similar good properties. At least for 2×2 games, the parameter-free concept of IBE permits explicit formulas for the predicted relative frequencies.[5]

[4] Refer to the online Appendix for the corrected Figure 11 of the Selten and Chmura paper and a table with significance levels of the comparison of the performance with and without loss aversion.

[5] IBE is parameter free since the double counting of losses is not based on parameter estimates, but on the idea that there are two reasons to avoid a forgone payoff. It is true that parameters for IBE have been estimated in previous papers on auctions (Axel Ockenfels and Selten 2005; Tibor Neugebauer and Selten 2006). However, these parameters do not relate to loss aversion in terms of lost money but rather to losing an auction. Loss aversion is relevant neither in Ockenfels and Selten (2005) nor in Neugebauer and Selten (2006) since in these auctions the pure strategy maximin is always zero. Thus, the parameters capture differences of feedback conditions and effects of social comparison. In settings in which losses (i.e., forgone money) and loss aversion were relevant, IBE was

In their comment, Brunner, Camerer, and Goeree (2011, 1038–39) state: "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." We do agree that nonparametric concepts like IBE have the advantage to serve as the basis of theoretical investigations just like NE. But we disagree with the second part of this statement. In fact, we hold exactly the opposite point of view. Parameter-free models allow clear predictions up front, and the empirical performance is easily measurable. This does not hold for parametric concepts, where parameters and thus "predictions" need to be adjusted to the data after collection. These parameter adjustments may help to "organize" the data, but what do we learn about underlying behavior in face of wide variation of parameter estimates even within the class of the simplest economic games?[6] A really fruitful contribution for future research would be the estimation of parameters out of the game structure or from subjects' cognitive abilities (e.g., memory).

REFERENCES

Avrahami, Judith, Werner Güth, and Yaakov Kareev. 2005. "Games of Competition in a Stochastic Environment." *Theory and Decision*, 59(4): 255–94.

Brunner, Christoph, Colin F. Camerer, and Jacob K. Goeree. 2011. "Stationary Concepts for 2 × 2 Experimental Games: Comment." *American Economic Review,* 101(2): 1029–40.

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.

Goerg, Sebastian J., and Reinhard Selten. 2009. "Experimental Investigation of Stationary Concepts in Cyclic Duopoly Games." *Experimental Economics*, 12(3): 253–71.

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.

Neugebauer, Tibor, and Reinhard Selten. 2006. "Individual Behavior of First-Price Auctions: The Importance of Information Feedback in Computerized Experimental Markets." *Games and Economic Behavior*, 54(1): 183–204.

Ockenfels, Axel, and Reinhard Selten. 2005. "Impulse Balance Equilibrium and Feedback in First Price Auctions." *Games and Economic Behavior*, 51(1): 155–70.

Selten, Reinhard, and Thorsten Chmura. 2008. "Stationary Concepts for Experimental 2 × 2-Games." *American Economic Review*, 98(3): 938–66.

Selten, Reinhard, Klaus Abbink, and Ricarda Cox. 2005. "Learning Direction Theory and the Winner's Curse." *Experimental Economics*, 8(1): 5–20.

applied as a parameter-free concept with double counting of losses (see Selten, Klaus Abbink, and Ricarda Cox 2005; Judith Avrahami, Werner Güth, and Yaakov Kareev 2005; Selten and Chmura 2008; Goerg and Selten 2009).

[6] Recall that Brunner, Camerer, and Goeree (2011) compute four different estimates for each of the three parametric concepts. Based on the datasets of the 2 × 2 games, of the symmetric matching pennies game (with and without loss aversion) and of the asymmetric matching pennies games, the estimates of the QRE parameter varied from $\lambda = 0$ to $\lambda = 3.62$, the ones of ASE from $n = 1$ to $n = 12$, and the ones of PSE between $n = 1$ and $n = 6$.