Supporting Information for Sergey Gavrilets & Peter J. Richerson "Collective action and the internalization of social norms"

Contents

Some known results on "us vs. nature" and "us vs. them" games

Our models here build on some earlier studies of "us vs. nature" and "us vs. them" games which did not include punishment or the possibility of norm internalization. In particular, Refs. (1–3) studied a series of related evolutionary models in which biological fitness (fertility) was defined as

$$w = 1 + bP - cx.$$

Assuming that individual effort x is controlled genetically (and thus is changed only by mutation), the equilibrium effort in "us vs. nature" games is predicted to be

$$x^* = \begin{cases} \frac{X_0}{n}\left(\sqrt{R}-1\right), & \text{if } R > 1, \\ 0, & \text{if } R < 1. \end{cases}$$

Here $R = \frac{b}{cX_0}$. That is, the effort is positive only if the benefit-to-cost ratio R is sufficiently large.

In "us vs. them" games, the equilibrium effort is predicted to be

$$x^* = \frac{1+b}{nc},$$

and thus is always positive. In both games, equilibrium efforts increase with benefit b and decrease with the group size n and cost parameters c (and X_0).

Numerical simulations

Dynamic variables. In our numerical models, each individual is described by three variables: the internalization trait η (whose value does not change during the individual life-time), the production effort x, and the punishment effort y. The latter two traits can change during the life-time if the individual updates them after collective actions. Trait η can take any value between 0 and 1 whereas traits x and y can take values only equal to 0 and 1.

Parameters. In numerical simulations we considered all possible combinations of the following exogenous parameters: benefit of collective action $b = 1, 2, 4$; normative values $v_x, v_y = 0, 0.5, 1.0$; group size $n = 8, 16, 24$; half-effort in "us vs. nature games" $X_0 = 0.25n, 0.5n, 0.75n$; cost of punishing $\delta = 0.25n/8, 0.5n/8, 1.0n/8$, cost of being punished $\kappa = K\delta$ with relative strength of punishment $K = 2, 3, 4$, and cost of monitoring $c_{mon} = 0.05n/8$. Note that the above implies that the costs of punishing, of being punished, and of monitoring group-mates increase linearly with group size n. Parameters that did not change are: number of groups $G = 500$, cost of collective action $c = 1$, the cost of optimization and internalization $c_{opt} = c_{int} = 0.05$, and the number of rounds per generation $Q = 40$. Genetic mutations occurred with probability $\mu = 0.0001$ per offspring, mutational effects were taken from a truncated normal distribution with zero mean and standard deviation $\sigma = 0.1$, Behavioral strategies x and y were updated independently after each collective action with probability $\nu = 0.25$. We ran simulations for 10,000 generations, 10 times each, for each parameter combination. The data reported are based on the last 10 rounds of the last 500 generations. The program was implemented in *Matlab*.

Initial conditions. Our simulations start with the internalization trait of each individual chosen randomly and independently from a uniform distribution on $[0, 0.05]$. The initial strategies (x, y) of individuals at the beginning of each generation (i.e., before round 1) are assigned randomly and independently with equal probabilities. This reflects the assumption that young individuals are susceptible to some extent to cooperative values promoted by the society at the very beginning of their adult life. However, they quickly start using their own judgment in making decisions afterwards. Fig. S1 illustrates the dynamics of the total group efforts in production, $X = \sum x$, and punishment, $Y = \sum y$, within a single generation over the course of $Q = 40$ collective actions (CA). Initial individual efforts in production ($x_i = 0$ or 1) and punishment ($y_i = 0$ or 1) are assigned randomly with equal probabilities. After each CA, a fraction $q = 0.25$ of individuals update their strategies using myopic optimization with errors. In Fig.1a (describing a case with a large average norm internalization), the groups converge on a state with high production (X) and punishment (Y). In Fig.1b (describing a case with with a low average norm internalization), the groups converge on a state with low production (X) and punishment (Y). (Notice higher between-group variation in X than in Y in both graphs.)

Additional graphs

Tables 1 and 2 present results of the analysis of variance of the effects of different parameters on production effort x, punishment effort y, degree of internalization η, material payoff π, and standard deviation in η. These Tables show that the normative value of punishment v_y and the group size n (in "us vs. them" games) have the highest influence of the models' dynamics.

Figures S2-S13 provide more complete results of numerical simulations of "us vs. nature" and "us vs. them" games.

Figures S14-S16 represent a sample of the data in Figures S2-S10 to illustrate better the effects of the half-effort parameter X_0 measuring the difficulty of the "us vs. nature" collective action.

Note that Figures S2-S16 are different from those in the main text in that they have an additional parameter K (relative strength of punishment) that varies between different runs. *Correspondingly, the color of bars has a different meaning from that in the main text.*

Additional discussion

Strategy revision protocol. Besides selection, a crucial factor controlling evolutionary dynamics is the strategy revision protocol (4). One common approach in studies of social behavior is to assume that individual strategies are controlled genetically and are changed only by random mutation (e.g., Ref. (5)). In this case, individuals are "stuck" with their strategies for life and any possibility of a free will or rational choice is absent. Another common approach is to assume that individuals rationally copy group-mates with higher payoffs (e.g., Ref. (6)). Applying this selective imitation approach to our models would allow maintenance of cooperation but not its emergence. This is because when common, free-riders in a group always have higher fitness than cooperators, leading the latter strategy to disappear. [Punishment and cooperation can invade by selective imitation if some individuals can remove themselves from collective actions and secure payoffs that are independent from other players' activities (7). We did not consider such loosely connected groups here.] Assuming payoff-based copying restricts individuals' abilities to make strategic decisions depending on the social environment. For example, the temptation to free-ride should clearly depend on the presence or absence of punishers in the group rather than on their current payoffs. Capturing these processes is possible within a bounded rationality approach used here. For example, in our model, individuals who do not internalize the norm will free-ride if there are no punishers in the group but may choose to cooperate if there are a sufficient number of punishers.

As acknowledged in the Methods section of the main text, we assume that each group members has complete information about the total contribution of its group and the total number of punishers among its group-mates. This assumption is not very realistic, especially for large groups. Future work might fruitfully explore relaxing this assumption to more boundedly-rational levels.

References

[1] Gavrilets, S. & Fortunato, L. A solution to the collective action problem in between-group conflict with within-group inequality. *Nature Communications* **5**, article 3526 (doi:10.1038/ncomms4526) (2014).

[2] Gavrilets, S. Collective action and the collaborative brain. *Royal Society Interface* **12**, article 20141067 (2015).

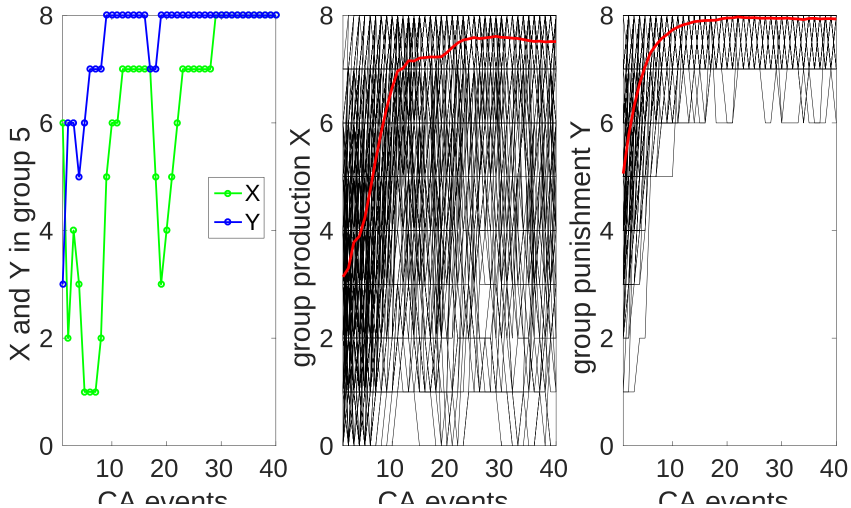
[3] Gavrilets, S. Collective action problem in heterogeneous groups. *Philosophical Transactions of the Royal Society London B* **370**, 20150016 (2015).

[4] Sandholm, W. H. *Population Games and Evolutionary Dynamics* (MIT Press, Cambridge, Massachusetts, 2010).

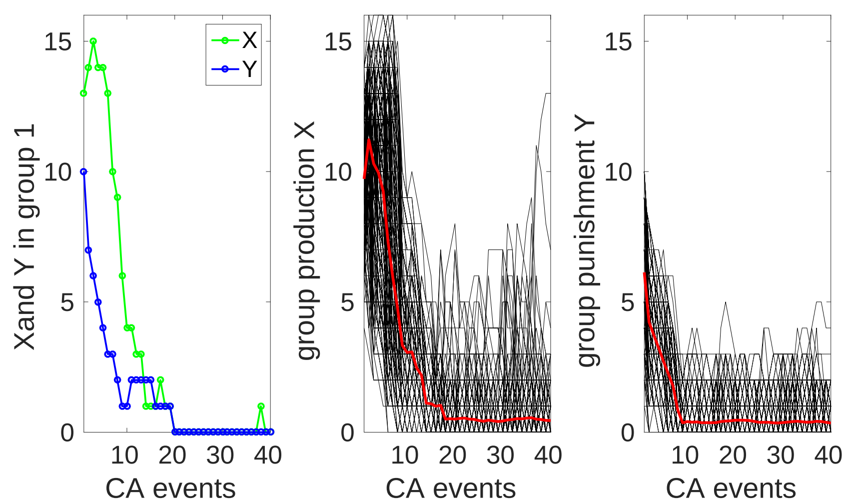
[5] Eldakar, O. T., Farrell, D. L. & Wilson, D. S. Selfish punishment: altruism can be maintained by competition among cheaters. *Journal of Theoretical Biology* **249**, 198–205 (2007).

[6] Boyd, R., Gintis, H., Bowles, S. & Richerson, P. The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences USA* **100**, 3531–3535 (2003).

[7] Hauert, C., Traulsen, A., Brandt, H., Nowak, M. A. & Sigmund, K. Via freedom to coercion: The emergence of costly punishment. *Science* **316**, 1905–1907 (2007).



(a)



(b)

Figure S1: The dynamics of collective action in a single generation. (a): "Us vs. nature" game when norm internalization does evolve (group size $n = 8$). (b): "Us vs. them" game when norm internalization does not evolve (group size $n = 16$). Left graphs: the dynamics of the total group effort in production $X = \sum x$ and in punishment $Y = \sum y$ in a single group. Center and right graphs: the dynamics of X and Y in the whole set of $G = 500$ groups. $Q = 40$ collective action events. Thin black lines: X and Y values for individual groups; thick red lines: the average across all G groups.

Table S1: Percentage of variation in different variables explained by different parameters in "us versus nature" games with $b = 4$ (analysis of variance with linear effects only).

Parameters	Variables				
	x	y	η	w	σ_η
n	-0.01	-0.12	-0.05	-0.05	+0.14
v_x	0.0	0.0	-0.01	0.0	-0.02
v_y	+0.43	+0.29	+0.18	+0.29	+0.19
H	+0.02	+0.02	+0.03	-0.22	+0.01
δ	+0.03	0.0	0.0	+0.02	+0.01
K	+0.03	0.0	0.0	+0.03	0.0
error	0.48	0.55	0.71	0.40	0.62

The absolute values shown give the percentage of the variation explained by a given parameter. The values smaller than 0.5% are shown as zeros (even if the corresponding effects are statistically significant). The sign shows whether increasing the corresponding parameter increases $(+)$ or decreases $(-)$ the corresponding variable statistically significantly. The absence of a sign means the corresponding effect (estimated using linear regression) is not significantly different from zero (at $P = 0.05$). x, y, η, w and σ_η are the production effort, the punishment effort, the internalization trait, the material payoff, and the standard deviation in η, respectively; H is the proportion of group members necessary for achieve half-effort, $H = X_0/n$.

Table S2: Percentage of variation in different variables explained by different parameters in "us versus them" games with $b = 1$ (analysis of variance with linear effects only).

Parameters	Variables				
	x	y	η	w	σ_η
n	-0.25	-0.39	-0.42	+0.09	+0.11
v_x	-0.00	-0.00	-0.01	+0.01	-0.01
v_y	+0.38	+0.29	+0.13	-0.44	+0.19
δ	+0.01	-0.01	+0.00	-0.08	0.00
K	+0.02	0.00	0.00	-0.04	0.00
error	0.34	0.31	0.44	0.35	0.70

The absolute values shown give the percentage of the variation explained by a given parameter. The values smaller than 0.5% are shown as zeros (even if the corresponding effects are statistically significant). The sign shows whether increasing the corresponding parameter increases $(+)$ or decreases $(-)$ the corresponding variable statistically significantly. The absence of a sign means the corresponding effect (estimated using linear regression) is not significantly different from zero (at $P = 0.05$). x, y, η, w and σ_η are the production effort, the punishment effort, the internalization trait, the material payoff, and the standard deviation in η, respectively.

Us versus nature games with $X_0 = n/4$ (half-effort is achieved with 25% participation in CA)

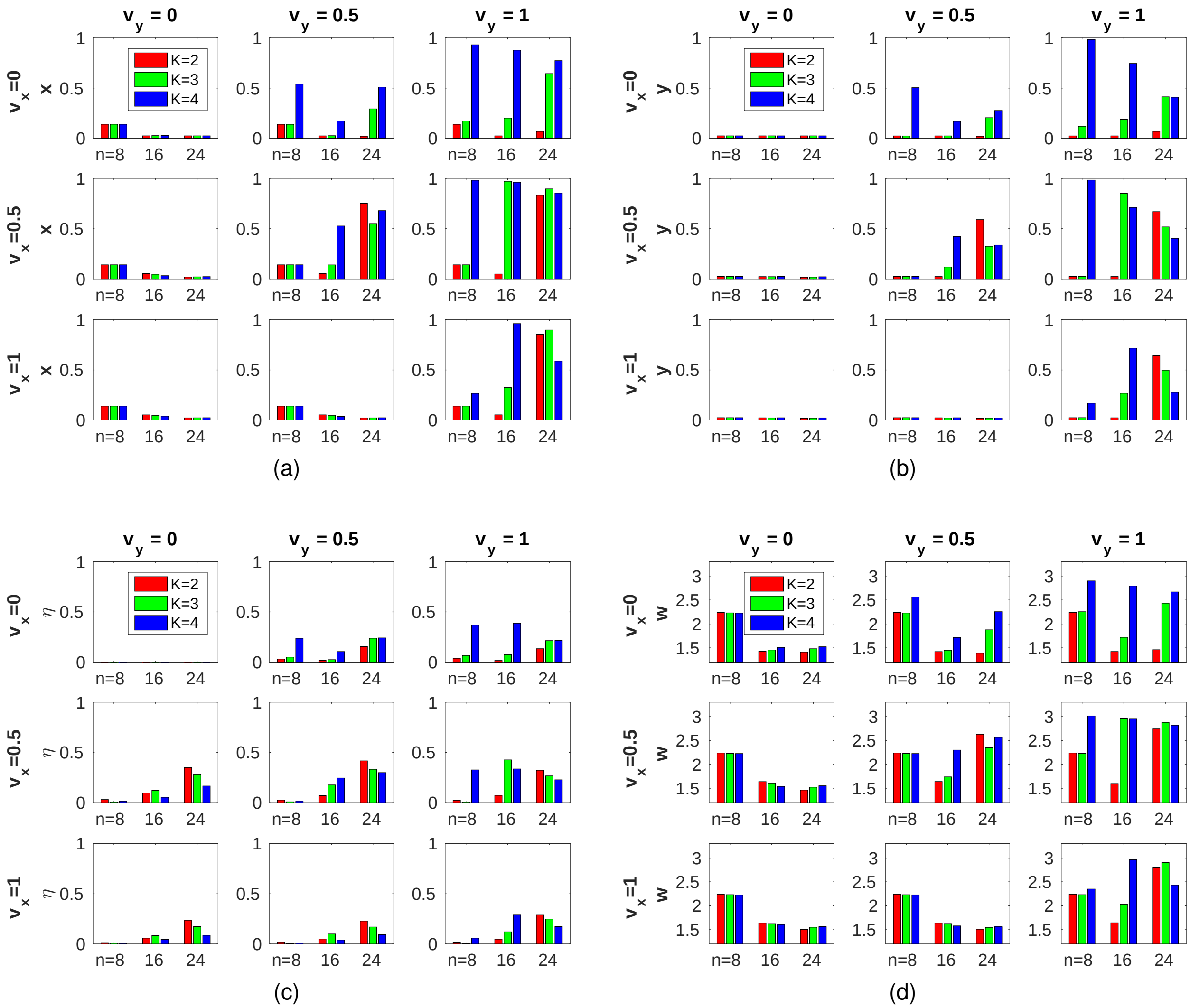


Figure S2: "Us versus nature" games: (a) efforts x, (b) punishment y, (c) internalization η, and (d) fitness w for different group sizes n, relative strength of punishment K, and normative values of production v_x and punishment v_y. $X_0 = n/4, \delta = 0.25, b = 4$. Shown are averages based on 10 runs for each parameter combination.

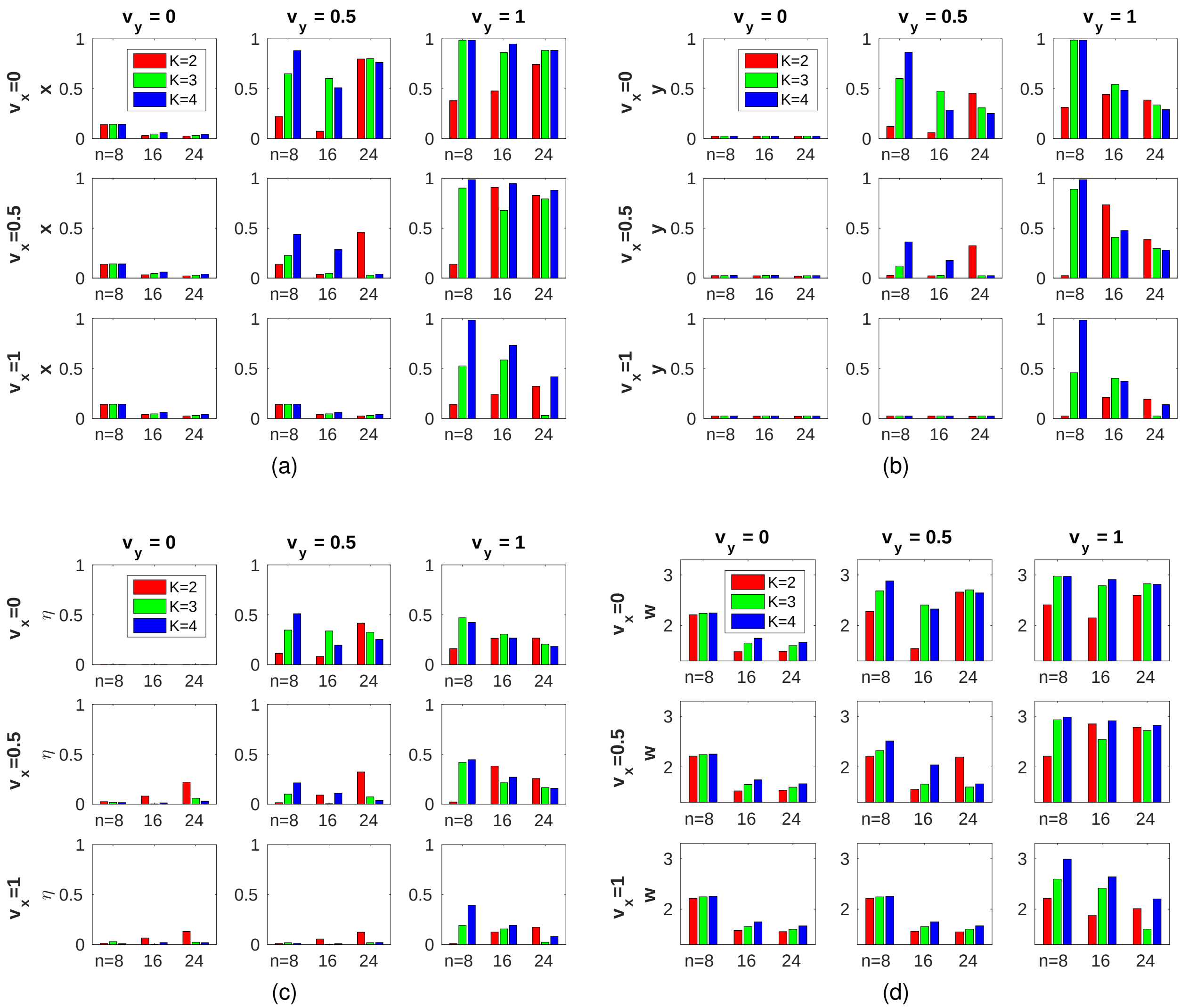


Figure S3: "Us versus nature" games: (a) efforts x, (b) punishment y, (c) internalization η, and (d) fitness w for different group sizes n, relative strength of punishment K, and normative values of production v_x and punishment v_y. $X_0 = n/4, \delta = 0.5, b = 4$. Shown are averages based on 10 runs for each parameter combination.

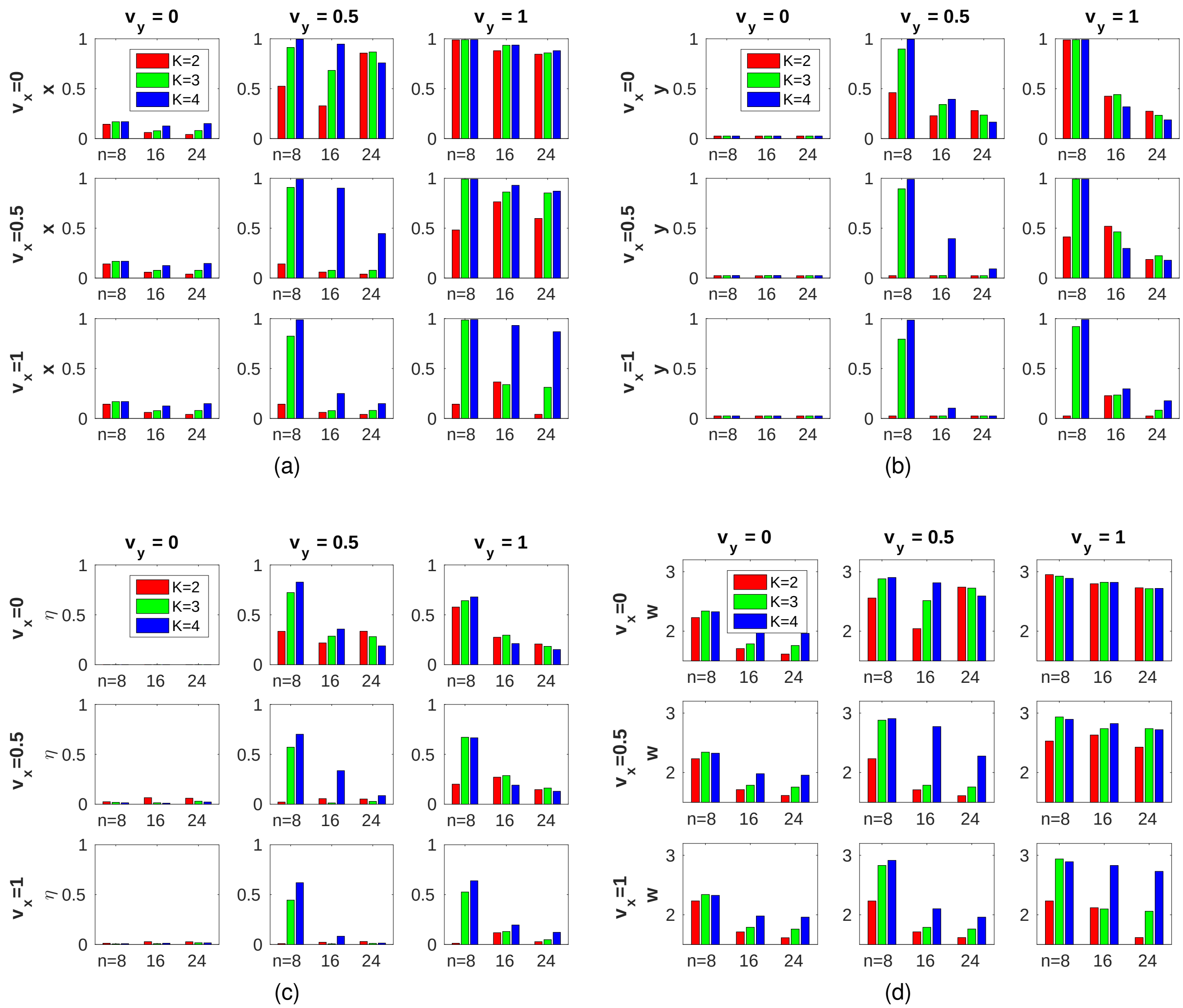


Figure S4: “Us versus nature” games: (a) efforts x, (b) punishment y, (c) internalization η, and (d) fitness w for different group sizes n, relative strength of punishment K, and normative values of production v_x and punishment v_y. $X_0 = n/4, \delta = 1.0, b = 4$. Shown are averages based on 10 runs for each parameter combination.

Us versus nature games with $X_0 = n/2$ (half-effort is achieved with 50% participation in CA)

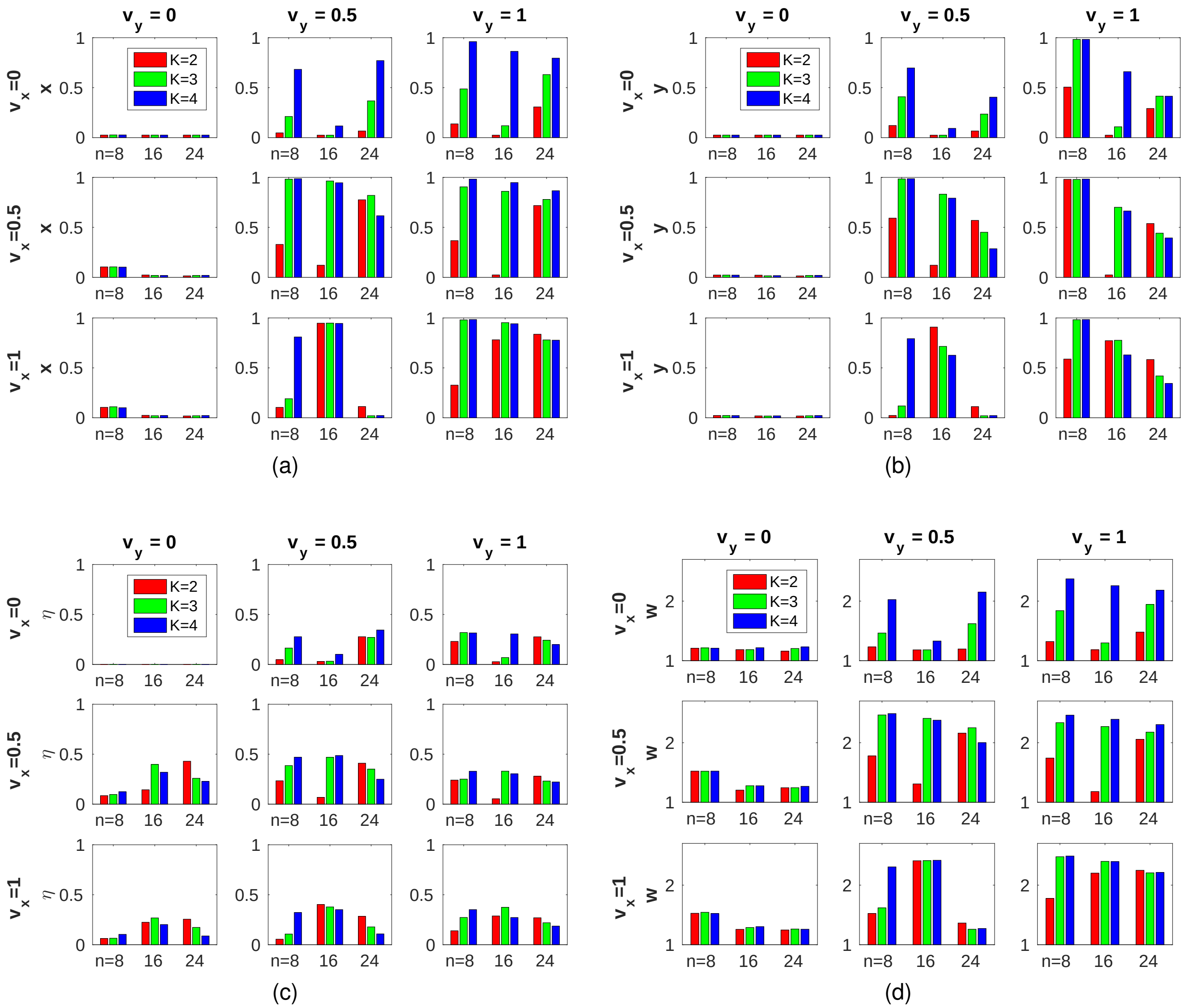


Figure S5: "Us versus nature" games: (a) efforts x, (b) punishment y, (c) internalization η, and (d) fitness w for different group sizes n, relative strength of punishment K, and normative values of production v_x and punishment v_y. $X_0 = n/2, \delta = 0.25, b = 4$. Shown are averages based on 10 runs for each parameter combination.

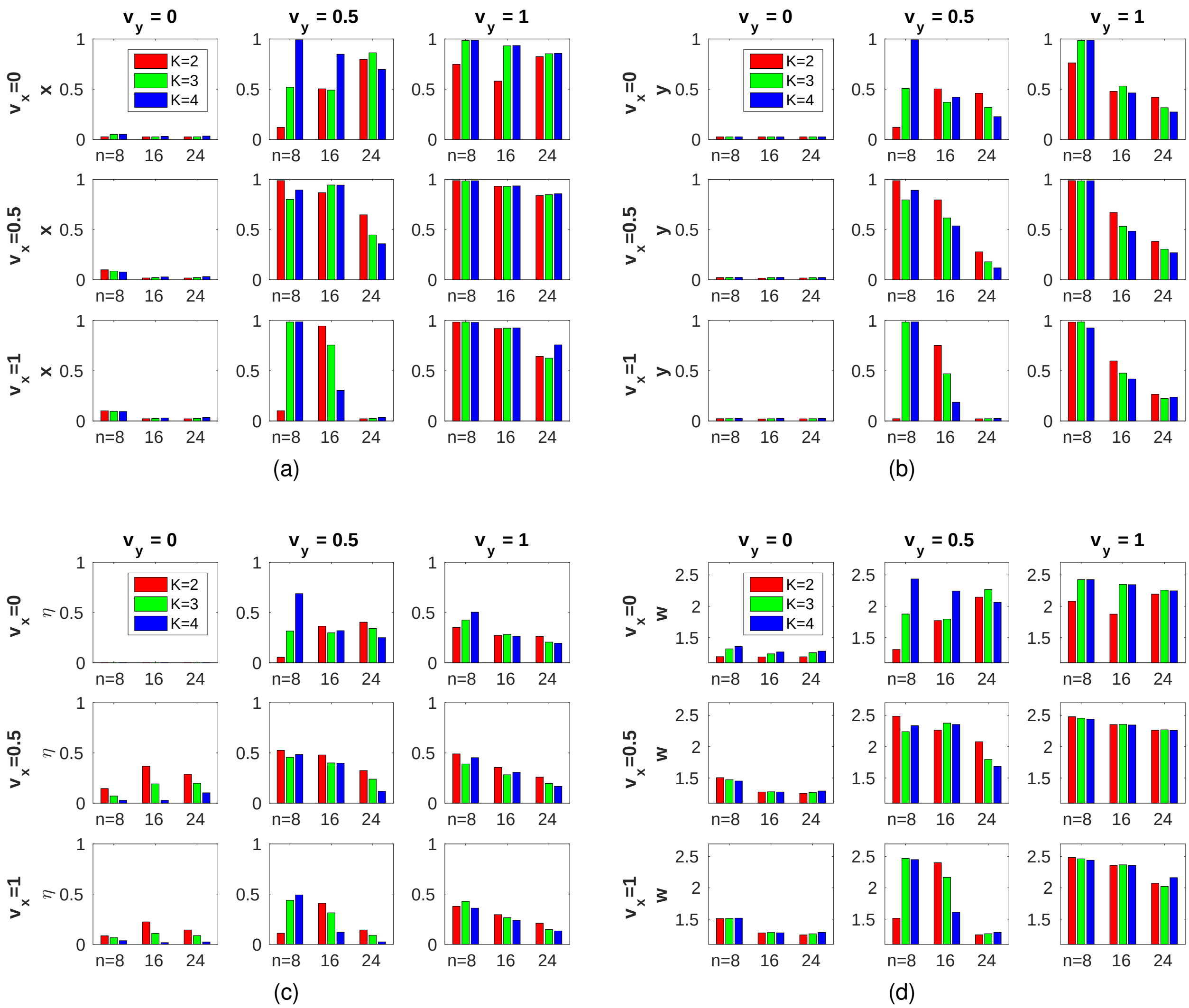


Figure S6: “Us versus nature” games: (a) efforts x, (b) punishment y, (c) internalization η, and (d) fitness w for different group sizes n, relative strength of punishment K, and normative values of production v_x and punishment v_y. $X_0 = n/2, \delta = 0.5, b = 4$. Shown are averages based on 10 runs for each parameter combination.

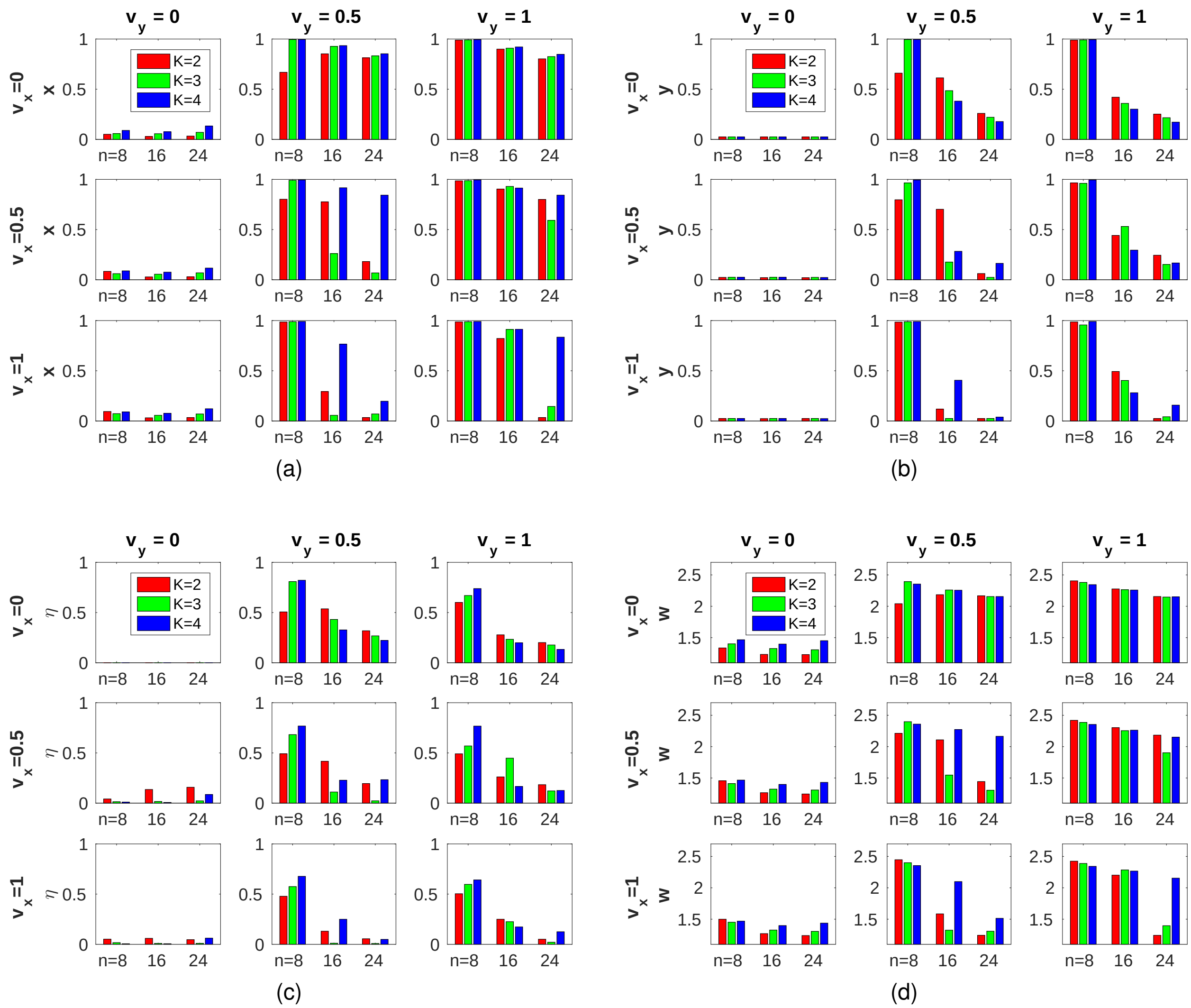


Figure S7: "Us versus nature" games: (a) efforts x, (b) punishment y, (c) internalization η, and (d) fitness w for different group sizes n, relative strength of punishment K, and normative values of production v_x and punishment v_y. $X_0 = n/2, \delta = 1.0, b = 4$. Shown are averages based on 10 runs for each parameter combination.

Us versus nature games with $X_0 = 3n/4$ (half-effort is achieved with 75% participation in CA)

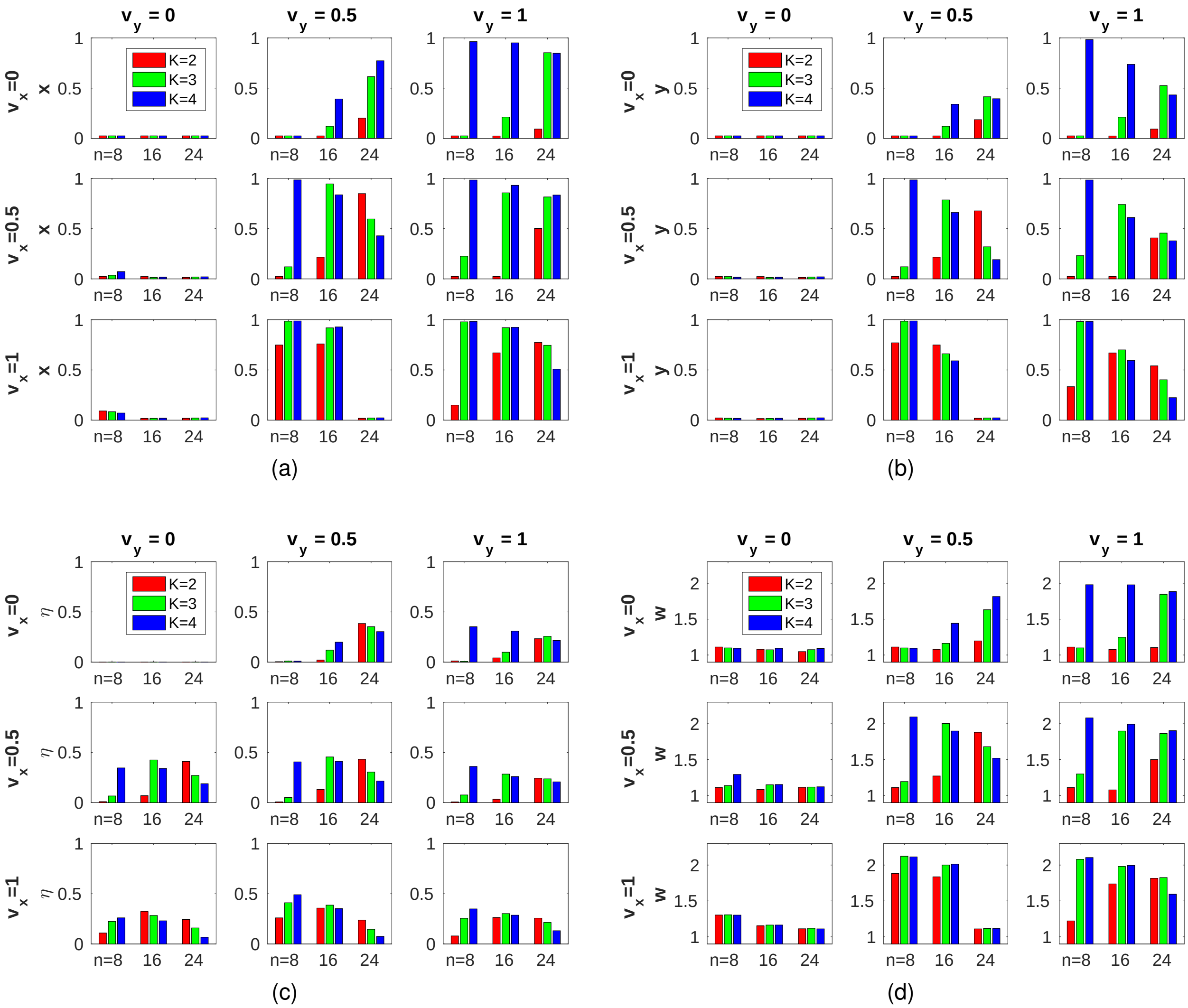


Figure S8: "Us versus nature" games: (a) efforts x, (b) punishment y, (c) internalization η, and (d) fitness w for different group sizes n, relative strength of punishment K, and normative values of production v_x and punishment v_y. $X_0 = 3n/4, \delta = 0.25, b = 4$. Shown are averages based on 10 runs for each parameter combination.

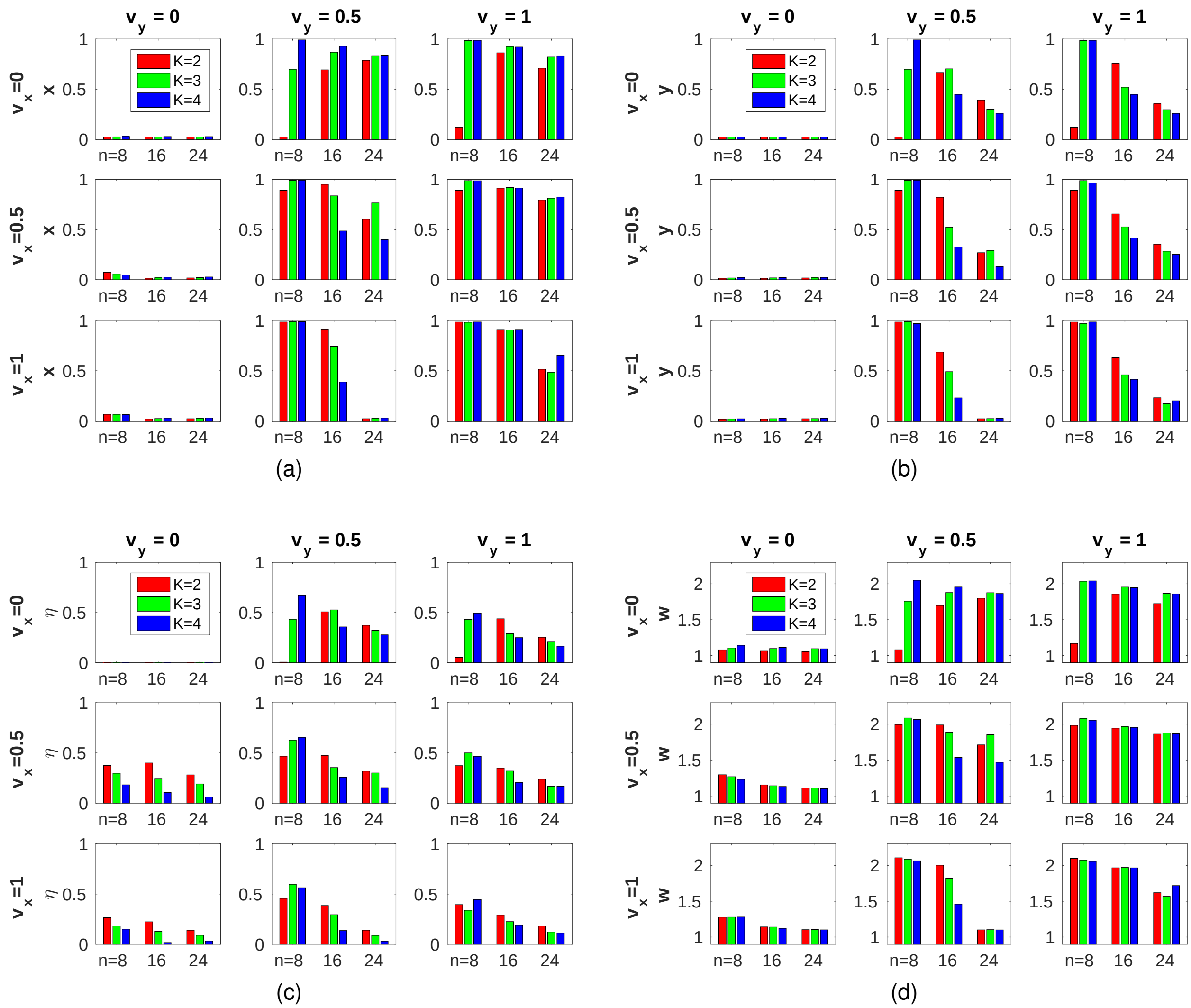


Figure S9: "Us versus nature" games: (a) efforts x, (b) punishment y, (c) internalization η, and (d) fitness w for different group sizes n, relative strength of punishment K, and normative values of production v_x and punishment v_y. $X_0 = 3n/4, \delta = 0.5, b = 4$. Shown are averages based on 10 runs for each parameter combination.

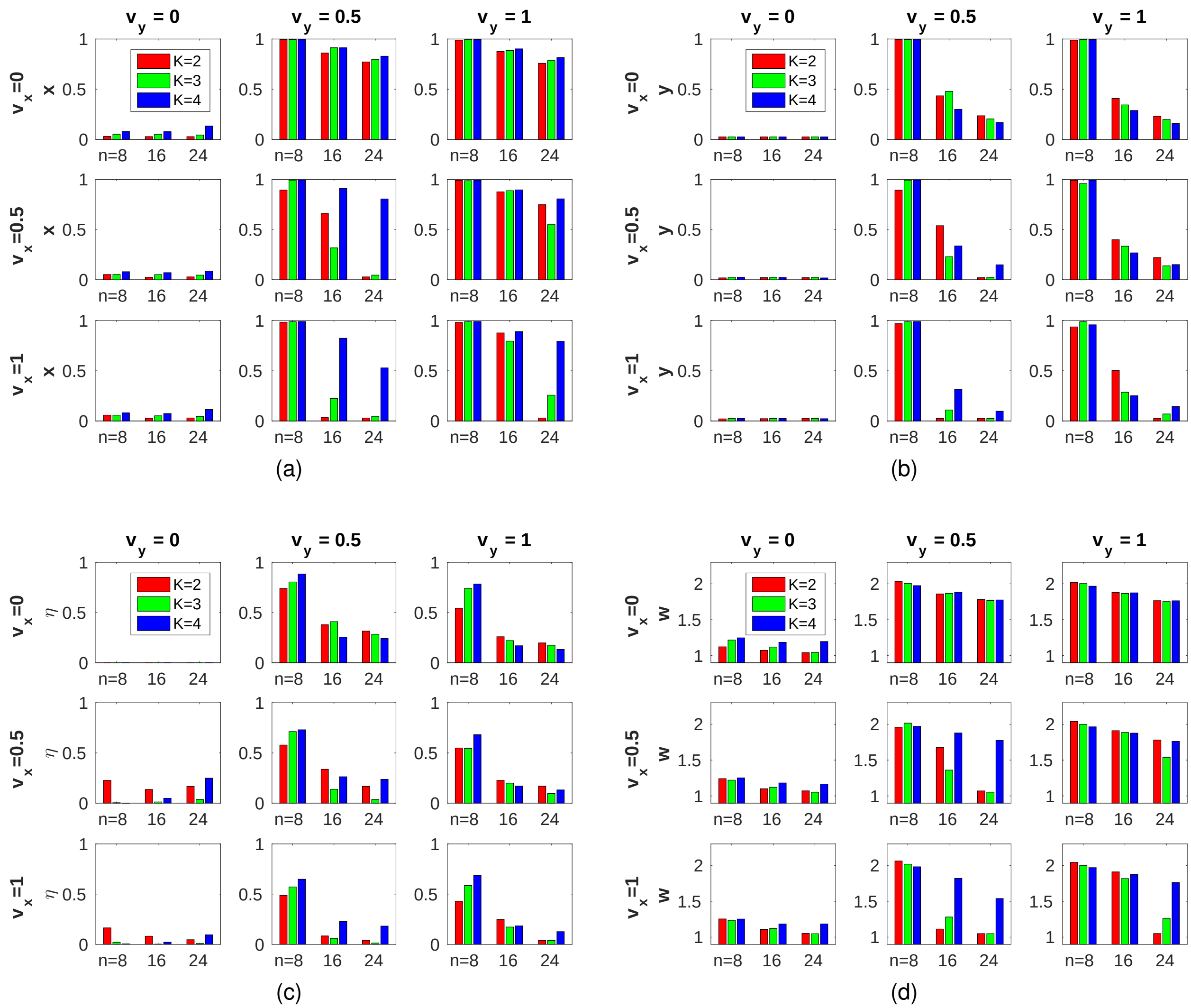


Figure S10: “Us versus nature” games: (a) efforts x, (b) punishment y, (c) internalization η, and (d) fitness w for different group sizes n, relative strength of punishment K, and normative values of production v_x and punishment v_y. $X_0 = 3n/4, \delta = 1.0, b = 4$. Shown are averages based on 10 runs for each parameter combination.

"Us vs. them" games

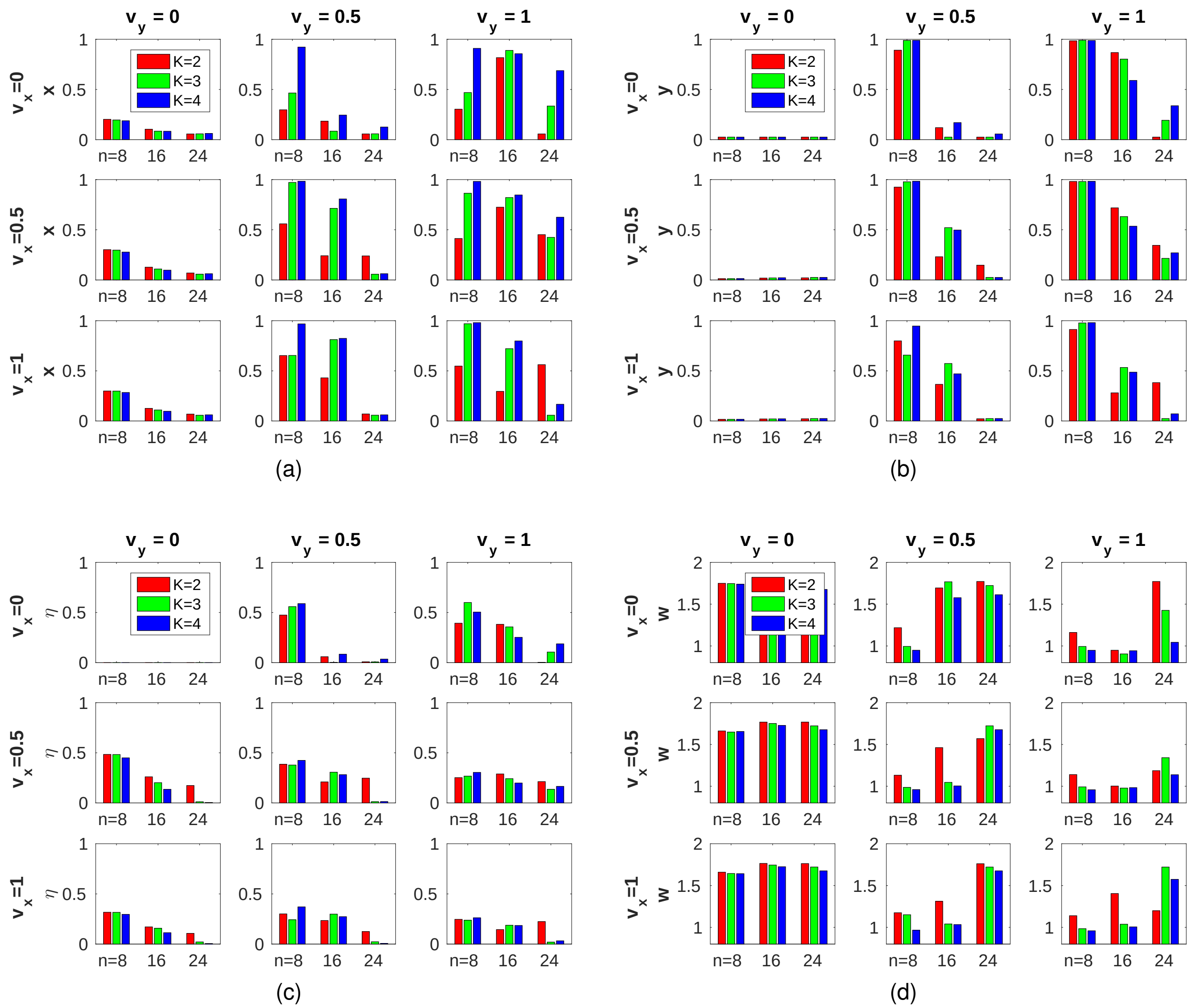


Figure S11: "Us versus them" games: (a) efforts x, (b) punishment y, (c) internalization η, and (d) fitness w for different group sizes n, relative strength of punishment K, and normative values of production v_x and punishment v_y. $\delta = 0.25, b = 1$. Shown are averages based on 10 runs for each parameter combination.

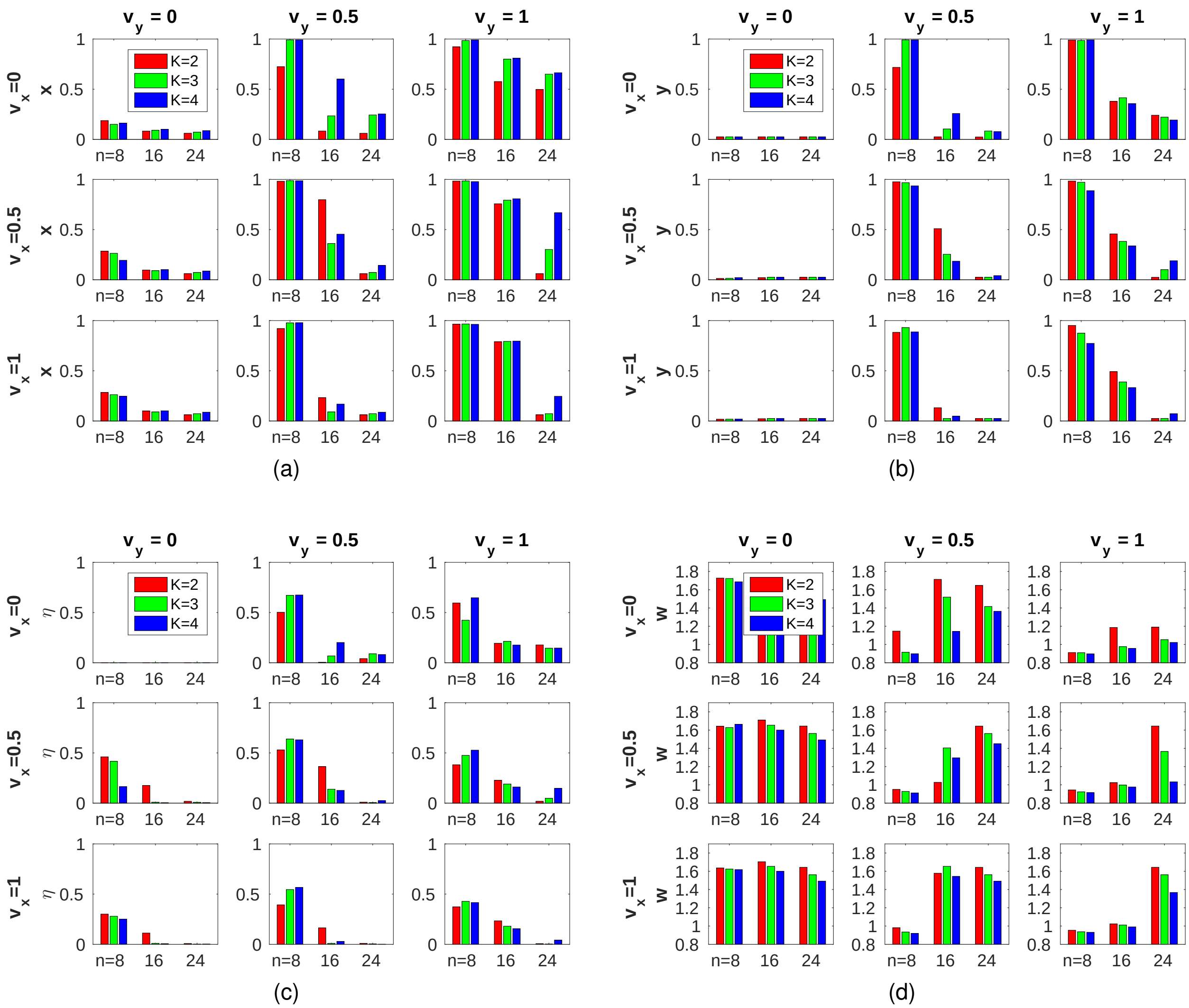


Figure S12: "Us versus them" games: (a) efforts x, (b) punishment y, (c) internalization η, and (d) fitness w for different group sizes n, relative strength of punishment K, and normative values of production v_x and punishment v_y. $\delta = 0.5, b = 1$. Shown are averages based on 10 runs for each parameter combination.

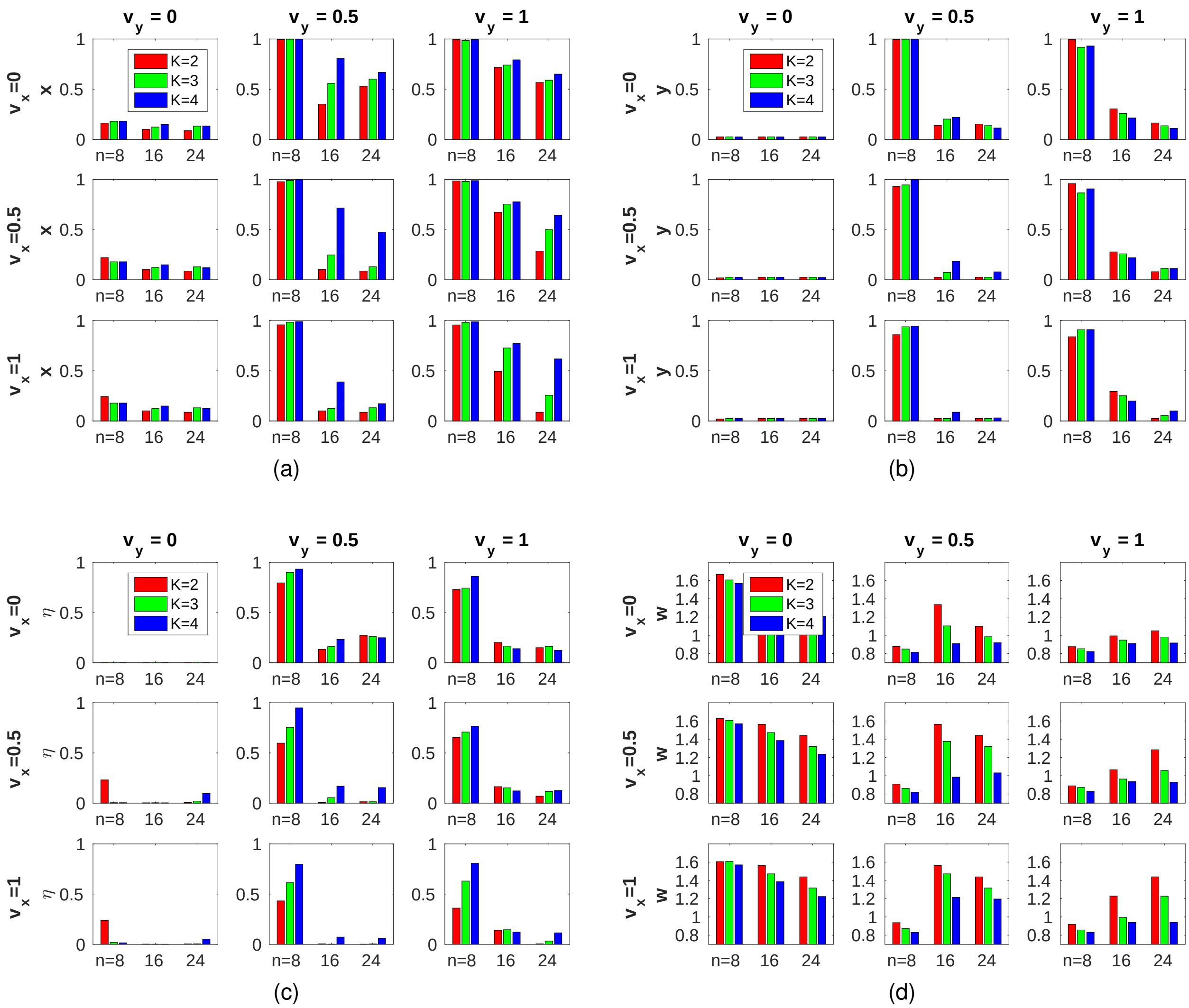


Figure S13: "Us versus them" games: (a) efforts x, (b) punishment y, (c) internalization η, and (d) fitness w for different group sizes n, relative strength of punishment K, and normative values of production v_x and punishment v_y. $\delta = 1.0, b = 1$. Shown are averages based on 10 runs for each parameter combination.

Effects of X_0 ("difficulty" of the task)

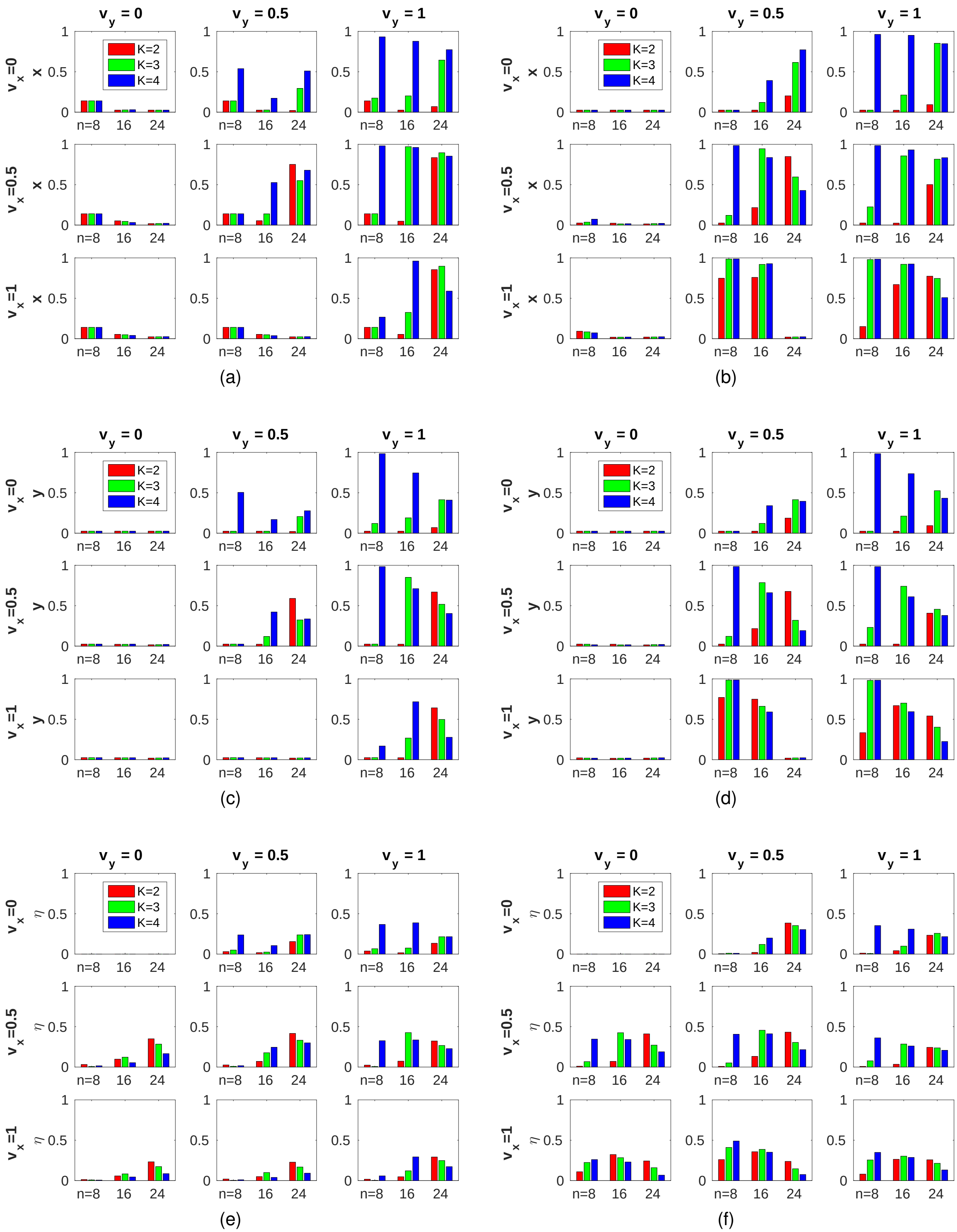


Figure S14: "Us vs. nature" $X_0 = n/4$ (easy task, left) vs. $X_0 = 3n/4$ (difficult task, right) at $\delta = 0.25$ (low cost of punishment) and $b = 4$.

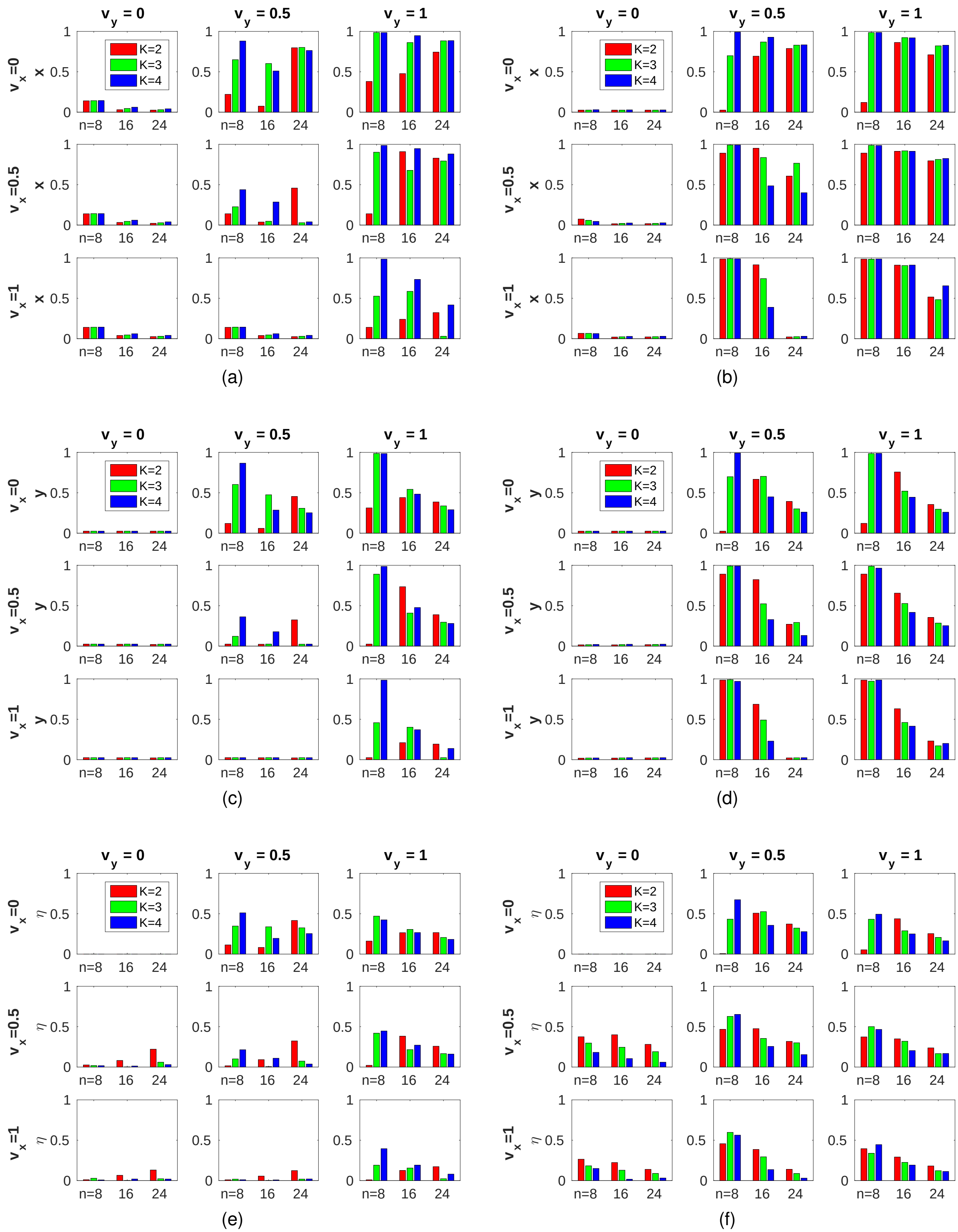


Figure S15: "Us vs. nature" $X_0 = n/4$ (easy task, left)vs. $X_0 = 3n/4$ (difficult task, right) at $\delta = 0.5$ (intermediate cost of punishment) and $b = 4$.

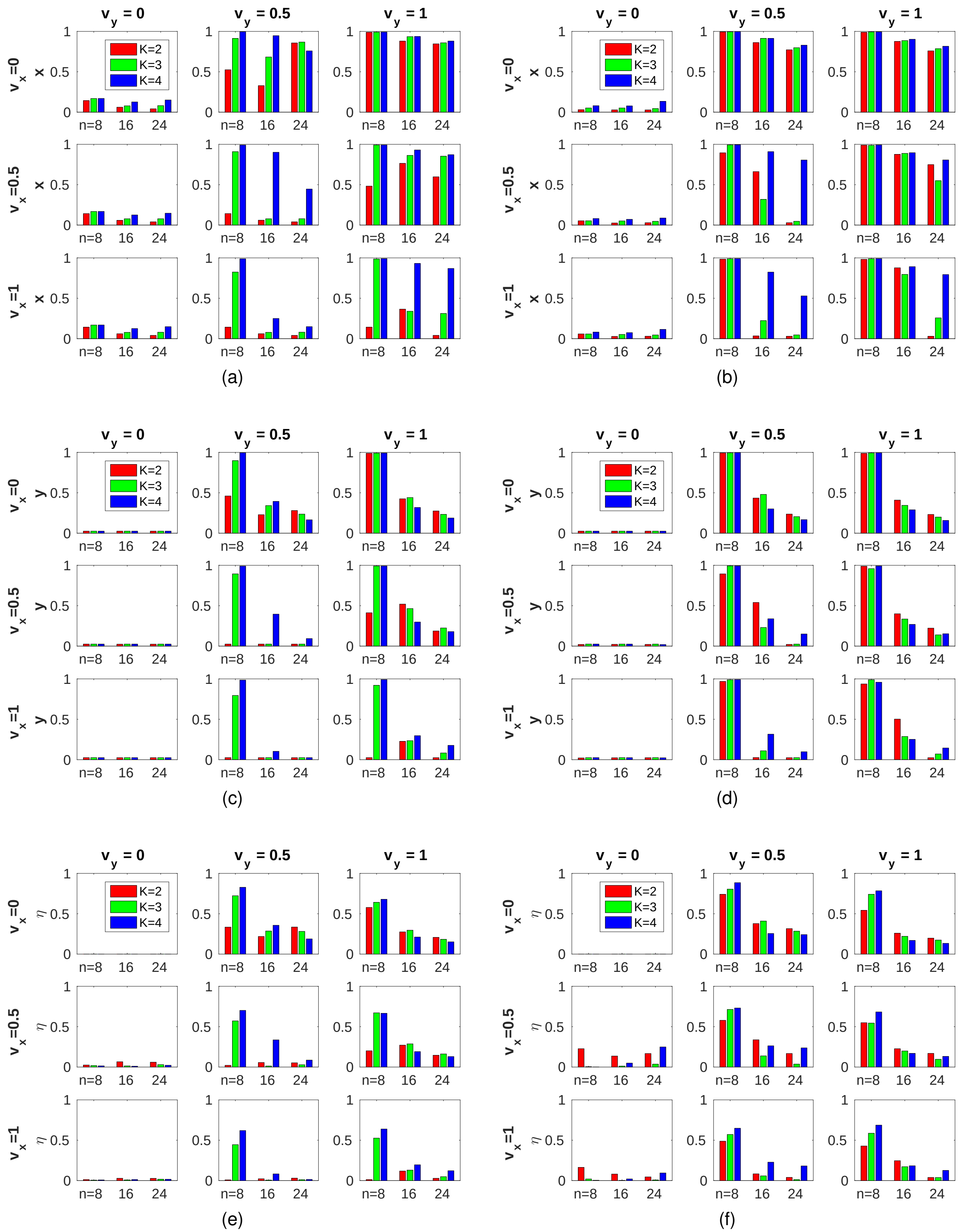


Figure S16: "Us vs. nature" $X_0 = n/4$ (easy task, left) vs. $X_0 = 3n/4$ (difficult task, right) at $\delta = 1$ (high cost of punishment) and $b = 4$.