Supplementary Methods for “Fish can infer social rank by observation alone”.

November 10, 2006

1 Subjects

The *Astatotilapia burtoni* used in this study were bred from wild-caught stock and are maintained at Stanford University under laboratory conditions that simulate those of their natural environment in Lake Tanganyika, Africa : pH 7.8-8.2, temperature 29 °C, 12h:12h light:dark cycle with full-spectrum illumination (Duralight 30 W; Bob Corey Associates, Merrick, NY, USA). Gravel and terracotta potshards provided visual isolation, allowing dominant males to establish and maintain territories, integral components of their reproductive and social behaviour. Fish were fed ad libitum once daily at 09:00-09:30h with cichlid formula pellets and flakes (Aquadine, Healdsburg, CA, USA). All work is performed in compliance with the animal care and use guidelines at Stanford University (protocol #3110)

2 Bystander Training on Fights Between Rivals

The two separated central males in each tank were surrounded by A through E arranged in separate units (Figure 1). Bystanders observed regular pair-wise fights A+B-, B+C-, C+D-, and D+E- between the stimuli fish, implying the hierarchy A>B>C>D>E. Bystanders were trained for 7 minutes per fight and saw two fights a day: one in the morning after feeding and one in the evening (a daily schedule is provided in Table I). The order in which these fights were shown to the bystanders was carefully counterbalanced so that the temporal order in which the fights were presented did not result in certain fights being first or last for all fish. In fact, for any fight at some point in the training schedule in which an AE or BD rival lost for some bystander “E1”, another bystander “E2” saw the other rival lose at the same point in the schedule. Five such schedule days are detailed in the top half of Supplementary Figure 1.

After two days of training bystanders were switched to the other side of the divided central compartment. Switching was always performed immediately following the evening fight on the second day. Once switched, bystanders again watched two days of fights as before, but now featuring the other half of the rival fish pairs (e.g. if bystander E1 saw A+B- and B+C- fights for the first two days, it would now see C+D- and D+E-, and E2 which had saw C+D- and D+E- the first two days would now see A+B- and B+C-). On the fifth day observers were shown all four pairs of rivals (two in the morning and two in the evening), with the switch occurring immediately after the two morning fights, allowing fish 2h to habituate to the move. On days six through ten the same schedule was repeated, and on day eleven the “Day 5” schedule was repeated. Prior to a fight, clear barriers between rivals were replaced with opaque ones. Fish were then allowed to habituate to these new conditions for 30 minutes.

When not being trained, bystanders were separated from each other by a clear barrier and housed with females to provide a standard social environment. Similarly, rivals were housed with females and allowed visual access to other rival males between fights. At least 30 min before fights, however, females were removed from the rival units and opaque barriers were added between all males. Opaque barriers separated bystanders from rivals at all times – except during scheduled fights, when the opaque barrier between the bystander fish and one unit containing the two fighting rivals was removed.

Table I shows a 5-day block of training for males in the hierarchy condition. Note that as a result if the spatial counterbalancing (with A and E in switched locations for half the bystanders, controlling for any possible spatial value transfer from the A and E end anchors to B and D), training for males with A-E arranged around them counter-clockwise E,B,C,D,A required a modified schedule ordering from that of males trained with A-E arranged around them counter-clockwise A,B,C,D,E. This modification was due to the division of the central unit into two bystander units and the related switching detailed above. It did not appear to effect the results in any way, and is detailed in the bottom half of Supplementary Table 1.

Bystanders were trained and data collected over three runs in two training tanks (N=4 bystanders per run). There were N=8 "hierarchy" bystanders trained on N=10 rivals, and N=4 "equated controls" trained on N=5 of the 10 previously used rivals as well as N=5 new rivals (see final section). This yielded 12 bystanders and 15 rivals overall (N=27). Only the N=8 "hierarchy" trained bystanders and their corresponding N=10 rivals are considered in the paper. The remaining fish in the "equated control" condition are discussed briefly in the final section below. The N=8 hierarchy bystanders were trained in pairs, with the first two pairs (N=4) each trained on five rivals (N=10 total). Rivals were then reused for the next two pairs (N=4), but with their order in the hierarchy inverted from their previous hierarchy position.

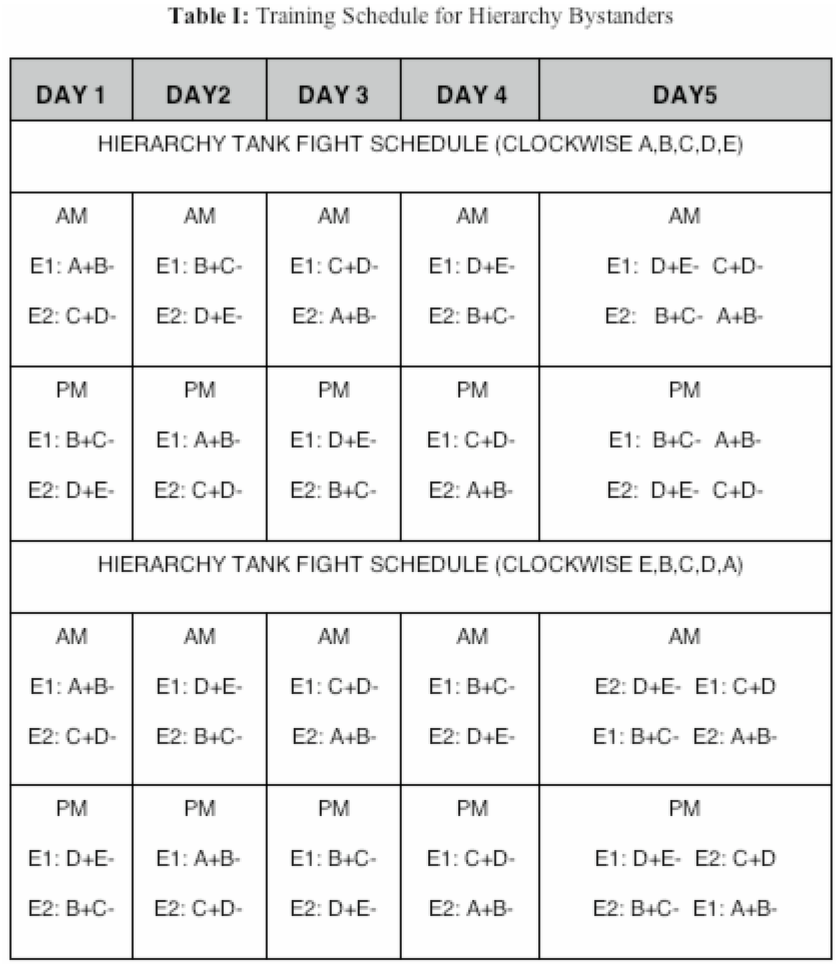


Table I: Training Schedule for Hierarchy Bystanders

DAY 1	DAY2	DAY 3	DAY 4	DAY5
HIERARCHY TANK FIGHT SCHEDULE (CLOCKWISE A,B,C,D,E)				
AM E1: A+B- E2: C+D-	AM E1: B+C- E2: D+E-	AM E1: C+D- E2: A+B-	AM E1: D+E- E2: B+C-	AM E1: D+E- C+D- E2: B+C- A+B-
PM E1: B+C- E2: D+E-	PM E1: A+B- E2: C+D-	PM E1: D+E- E2: B+C-	PM E1: C+D- E2: A+B-	PM E1: B+C- A+B- E2: D+E- C+D-
HIERARCHY TANK FIGHT SCHEDULE (CLOCKWISE E,B,C,D,A)				
AM E1: A+B- E2: C+D-	AM E1: D+E- E2: B+C-	AM E1: C+D- E2: A+B-	AM E1: B+C- E2: D+E-	AM E2: D+E- E1: C+D E1: B+C- E2: A+B-
PM E1: D+E- E2: B+C-	PM E1: A+B- E2: C+D-	PM E1: B+C- E2: D+E-	PM E1: C+D- E2: A+B-	PM E1: D+E- E2: C+D E2: B+C- E1: A+B-

3 Preference Testing

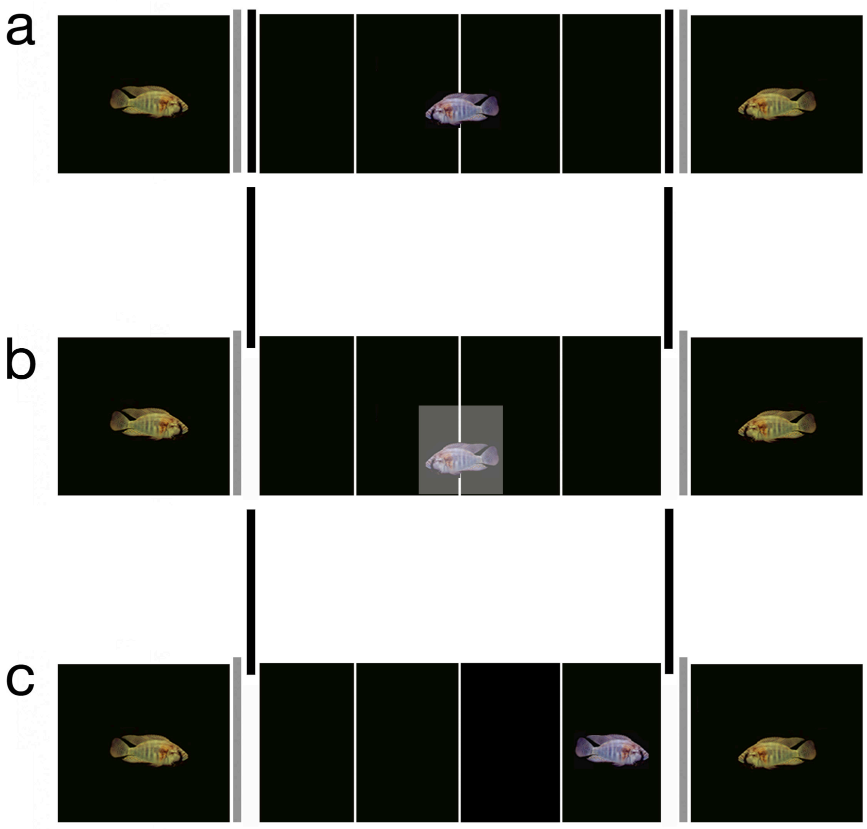
After bystanders were trained, preference for A vs. E and for B vs. D pairings was assessed using an approach/avoidance task. This task was run once in the training tank ("Familiar Context", with bystander choice area dimensions of $45.72l \times 22.86d \times 25.4h$ cm) and once in a tank designed for testing approach/avoidance preference ("Novel Context", with bystander choice area dimensions of $74l \times 37d \times 28h$ cm). Half of the bystanders were assessed in the training tank first while the rest were run in the test tank first, to control for possible order effects on fish discrimination. Similarly, whether fish were tested on the AE or BD pair first was counterbalanced, as was which sides of the test tank members of the AE or BD pairs were placed in as stimuli. Additionally rivals were always moved to the test tank first so they were never used in a test without having first been moved. They were allowed to acclimate for 30 minutes with two non-gravid females after being moved to the Novel Context or back to the Familiar Context. To control for preference towards a particular side of either tank, 10 minute baseline trials were collected, in which the tank was divided lengthwise into four quadrants and the amount of time spent by the fish in each quadrant was recorded (Suppl. Fig. 1a).

Following baseline, observers were placed in the center of the tank under a clear plastic box, ensuring equal exposure to stimuli males and an identical starting location between trials. After allowing the observer 1 minute to acclimate to the box, the opaque barriers at either end of the tank were removed simultaneously, exposing the AE or BD stimuli males positioned on opposite ends of the tank behind clear acrylic barriers (Suppl. Fig. 1b). After 10 minutes of this "forced viewing" condition, the box was removed, so that the bystander could swim freely about the tank (Suppl. Fig. 1c). First choice was recorded as the first movement into one of the far tank quadrants adjacent to a rival male. The bystander was then allowed a one-minute free swim before time data were collected. All novel context trials were recorded using a JVC digital video camera on a tripod 3m from the tank, while familiar context trials were recorded from above via a webcam suspended above the tank.

During the forced view, rival male activity was given a forced view score (FV) measuring aggressive behavior on a categorical scale from 0-3: (0) little or no movement, (1) movement but no eyebar activation or threatening behavior, (2) eyebar activation and less than 10 threatening actions, and (3) eyebar activation and more than 10 threatening actions. Logistic regression of first choice on the difference in FV between AE or BD in both contexts (FV) was non-significant (z=1.530, p=0.126). Removing two outliers (Cooks Distance $>$ 0.5 larger than the standard heuristic $4(n-k-1)$ threshold for possible outliers – in this case 0.133), left the regression non-significant (z=0.052, p=0.959). Linear regression of time spent on FV also failed to yield significant results (t=1.446, p=0.159). To make sure that any independent preference of the bystander for one side of the tank over another did not influence our results, we also regressed $Time\ spent$ on $Time\ spent\ during\ baseline$ (the same difference between time spent in the quadrants adjacent to each rival as in $Time\ spent$ but taken from the baseline observations). The regression was non-significant (t=0.109, p=0.914), showing that for these bystanders time spent in the tank when rivals were not visible did not predict time spent when they were.

4 Test Context Sensitivity

As the two contexts in which the bystanders were tested differed in overall dimension (with the "familiar" context being shorter and narrower than the "novel" context), and given that we are making claims about differences in bystander behaviour in different contexts, it is worth exploring the possibility of differences in tank sensitivity. If the two differently-sized tanks had different overall levels of sensitivity (e.g., the cost of choice being much greater in one than in the other)



Supplementary Figure 1 | Assessing Preference for Novel Pairings. Time spent by bystanders in each of four quadrants dividing the tank was recorded in three conditions lasting 10 minutes each. After introduction to the tank, a baseline (a) with opaque dividers visually isolating the bystander (coded here as blue although all fish in experiment were yellow) from the rival males (yellow) was taken for 10 minutes. This was followed by a 10 minute "forced view" during which the bystander was kept in place using a clear plastic box (b). Finally the box was removed and time spent recorded as the subject was allowed to swim freely about the tank (c). This protocol was repeated in both the familiar and the novel contexts.

we would expect this difference to have a global effect on the bystanders decisions between contexts. We thus checked for any large group differences in time spent near the different rivals by comparing across contexts using one-sample t permutation tests on the difference between contexts with 500,000 replicates of each comparison (as used previously in our analysis of the $Time\ Spent$ data and discussed further below). In no case did a difference between contexts (each $n = 8$) reach significance (near A: $p = 0.294$, near E: $p = 0.4$, near B: $p = 0.31$, near D: $p = 0.698$). This suggests that there are no detectable differences in time spent by bystanders next to the different rivals as a result of context at this sample size. Still, as we obviously cannot confirm the null hypothesis, the possibility of an effect too small to detect at this sample size remains. However, it is important to point out that these distances (45.72 cm in the familiar tank, 74 cm in the novel tank) are not large relative to the distances these fish swim in their normal habitat, differing by less than three bystander body lengths.

Finally, assuming there were a difference in sensitivity between contexts, it seems extremely unlikely that this difference would somehow manifest as a significant positive linear correlation in the Novel Context with C rival behavior during training. This is especially true given that we see the more graded response in the Novel Context, which since bystanders have to swim a greater distance between the end quadrants (in which they spend nearly all of their time during preference), if there were a difference it should encourage a less graded response – as bystanders would be less likely to swim the greater distance back and forth between rivals in the longer tank given the greater cost of doing so. Again, however, these distances are trivial for these fish to traverse and there is likely no difference in sensitivity between contexts as suggested by the tests above.

5 Dependence of Approach/Avoid Measures Between Contexts

Importantly, it would be undesirable to test the bystanders in the different contexts on different sets of rivals. When testing for differences in choice behaviour between rivals in the different contexts, testing the same bystander on the same rivals in both contexts controls for variance across different bystanders, rivals, and training runs. Such variance might otherwise account for differences across contexts beyond that due to the context manipulation in which we are interested. Using different rivals – and perhaps more importantly a different training run, which stochastically would have, for example, different C rival behavior – might artificially inflate the difference measured between the fits to the different contexts. We are interested in what happens when we vary context only.

Still, it is very appropriate to discuss the dependence of the bystander responses when considering our comparison of the regression slopes between contexts. Note that this dependence is irrelevant until the analyses are compared, as all testing up to this point was done separately for the different contexts.

There are two reasons for the approach/avoid measures on the bystanders to be dependent between contexts:

1. They are repeated measures on the same bystanders in different contexts. Even if the rivals were different in each context, the approach/avoid measures would be dependent in this regard.
2. The rivals were not different in the different contexts. Rather, the bystander was trained once on a set of 5 rivals and then chose between the same AE or BD rivals twice, with the choices made in two different contexts.

We address the concerns related to these dependencies separately. For the between context comparison the repeated measures on the bystanders are obviously an important issue, as the test for equality of slopes is equivalent to running a single ANCOVA with *Context* coded as a dummy variable, and data from both contexts are therefore used in one model. The interaction of *Context* with *C Rival Behavior* then tells us whether the difference between the slopes is statistically different from zero. We take this into account by using a repeated-measures ANCOVA, which accounts for the bystander dependencies across contexts.

Regarding the second source of dependency, in psychology is not unusual to compare the reactions of a subject to the same stimuli before and after some manipulation (in this case change of context). The primary dependency concern in such a design is subject learning or carry-over effects. We have controlled for such effects in the standard way by using a crossed design, counterbalancing the order in which the bystanders were tested in the different contexts. As any carry-over effects from such learning should be relatively symmetric, this crossed design is a sufficient control. Such crossing would necessarily disrupt any linear relationships due to carry-over effects, and were there any carry-over effects (symmetric or not), should result in distinct subpopulations within the contexts. We do not see such results, finding instead graded effects indicative of linear relationships with the independent variable.

6 Scoring and Inter-rater Reliability

All video recordings taken of the baseline and preference tasks in both the “Familiar” and “Novel” contexts were intentionally framed so that only the portion of the tank containing the bystander male was visible (i.e. so that the rival males were not visible). As the rival positions were regularly reversed in the counterbalancing mentioned above, the tapes randomized before scoring, and the bystanders weight/size/color matched, the scorer did not know which rival was on which side while scoring the video nor which bystander was being scored. The scorer was thus blind to which condition the bystander was in, and to where the higher and lower ranking rivals were located. As the subjects in this experiment are clearly also “blind”, the scoring was double-blind.

However, to be absolutely sure that there was no systematic bias on the part of the scorer, who had also run the experiment, we had a naïve student new to the experimental system and completely unaware of the experimental design or hypothesis rescore two samples of eight experimental trials (one sample from each context in total, including every bystander at least once). We then compared *Time Spent* as calculated from her rescoring to that of the original scoring using a standard test of inter-rater reliability for quantitative data: the intraclass correlation coefficient (two-way random effects model for absolute agreement with single measure reliability) [3, 6]. For the Familiar Context (N=8) this yielded an agreement score between the two raters of 0.99 with an associate p-value of $p = 2.88 \times 10^{-8}$. In the Novel Context (N=8) the agreement score was also 0.99 with an associated p-value of $p = 2.43 \times 10^{-7}$.

These results reveal several important feature of the scoring. First, scoring these trials is an unambiguous task, as an untrained undergraduate following simple instructions was able to produce results almost perfectly matching those of a more experienced scorer. Second, because it was a test of absolute agreement, not just consistency, the raters assigned essentially the same absolute score in addition to being almost perfectly correlated, indicating no detectable rater bias. In fact, overall the naïve student's *Time Spent* scores were very slightly higher (0.12 min) on average, suggesting that if anything, the original scoring was a bit more conservative than the naïve rescoring (the difference being miniscule). Finally, as for such ratings a score of 0.80 is considered "outstanding" [4], these scores indicate near perfect rater agreement, showing that the original scoring was without bias.

7 Resampling methods

Since we cannot adequately test that parametric assumptions (i.e. normality) hold at this sample size ($n = 8$), we have switched to a more appropriate resampling approach, using a permutation test for the one-sample t-test [1] on *Time Spent* against the null hypothesis that differences in time spent next to different rivals will not be systematically higher than "random" (i.e. centered at zero). Such tests have been recommended for small sample situations in ecology research [5] and more generally [7], as while they do not require the data to be normally distributed, they also do not suffer from the lack of power that standard nonparametric (e.g. ranking) tests typically display at small sample sizes, they are robust, and they can be made arbitrarily accurate via continued resampling [2].

The software we used to compare sample means is available free from the R Project for Statistical Computing (reference 30 in the manuscript) and found in the "Data Analysis and Graphics (DAAG)" package. Introductions to permutation testing and its advantages are readily available online [2].

8 Principal Component Analysis (PCA) and Dominance Score (DS)

For our behavioral data, the Kaiser-Meyer-Olkin Measure of Sampling Adequacy yielded a score of 0.83, and Bartlett's Test of Sphericity was very significant $p \ll 0.001$, confirming our data to be highly multicollinear and therefore well-suited to Principal Components Analysis (PCA). Dominance Score (DS) was calculated via PCA (using the PRINCOMP function in R), with variables $Chasing$, $Biting$, $Lateral\ Threat$, $Fleeing$, and $Eyebar\ Activation$ (as a percentage of fight time) as input variables. The first principal component extracted had an eigenvalue of 3.04, and indexed dominant/submissive behavior. The eigenvalues for the correlation matrix were 3.04, 0.80, 0.57, 0.34, and 0.26, with the first component corresponding to 60.17% of data variance. Each Dominance Score for each fish in each fight (DS) was then just the projection of the original data point onto the 1st eigenvector of the correlation matrix.

For each bystander, scores indexing combined rival behavior were calculated as follows. Let $D_{ij}^{(Winner)}(r)$ be the ith dominance score of rival $r \in \{B, C, D\}$ in which r was the winner and the fight was seen by bystander j. Let $D_{ij}^{(Loser)}(r)$ be equivalent except that j was the loser, and let $Z_X(x) = \frac{x-\mu_X}{\sigma_X}$ be the z-score of x in sample X. Then the "combined rival behavior" B_j for rival r seen by bystander j is:

$$B_j = \frac{1}{2}\left(Z_J \left| \sum_{i=1}^{W_r} D_{ij}^{(Winner)}(r) \right| + Z_J \left| \sum_{i=1}^{L_r} D_{ij}^{(Loser)}(r) \right| \right)$$

for all $j \in J$, where J is the sample of bystanders, and W_r is the total number of wins and L_r the total number of losses for rival r, and $|\cdot|$ the absolute value. We thus regressed each $(Time\ spent\ near\ B - Time\ spent\ near\ D)_j$ on B_j for all bystanders on rivals B, C, and D in each context to obtain the six results presented in the main text. It was decided prior to running the tests that the p-values would be corrected by multiplying each p-value by the number of tests before comparing at the standard $\alpha = 0.05$ level (Bonferroni correction). This rather conservative approach to type I error control was necessary due to the small number of data points ($n = 8$) being fit. To obtain the original p-values simply divide by 6.

9 Equated Controls

As a secondary check on our training procedure, we ran four additional bystanders ("equated controls") ($n = 4$) that were trained like the "hierarchy" bystanders except that all rivals won and lost equally (implying A=B=C=D=E). On day 5 type days "equated controls" were trained on fights implying either A<B<C>D>E (Day 5 schedule) or A>B>C<D<E (Day 5' schedule), ensuring that AE and BD stimuli maintained equal relative rank.

We anticipated that we would not see consistent choices in fish trained on a non-hierarchy. If, however, all four fish had demonstrated consistent preference for one side of the tank when trained on the four equated rivals, further study would have been necessary. However, these animals showed inconsistent preferences and we terminated this exploratory experiment.

References

[1] P. Good. *Permutation Tests.* Springer, New York, 2000.

[2] Hesterberg, T., Monaghan, S., Moore, D. S. Clipson, A & Epstein, R. *The Practice of Business Statistics*, chapter 18: Bootstrap Methods and Permutation Tests. W. H. Freeman and Co., 2002.

[3] D. C. Howell. *Statistical Methods for Psychology.* Duxbury, 2002.

[4] Landis, J. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics*, 33:153–174, 1977.

[5] Potvin, C. & Roff, D. A. Distribution-free and robust statistical methods: viable alternatives to parametric statistics. *Ecology*, 74(6):1617–1628, 1993.

[6] Shrout, P. E. & Fleiss, J. L. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 2:420–428, 1979.

[7] Tibshirani, R. & Efron, B. Statistical data in the computer age. *Science*, 253:390–395, 1991.