DO AGGRESSIVE DISPLAYS ESCALATE OR DEESCALATE CONFLICT? EVIDENCE FROM HUMAN CONTESTS

Abstract

Aggressive displays may have evolved as a substitute for outright combat (Lorenz, 1966; Skyrms, 2009; Enquist, 1985), i.e. to help discourage weaker contestants from futile conflict. We asked whether aggressive displays could discourage weaker humans from competing over resources. Our subjects knew that such displays were completely uninformative about their opponents strategy: they were instructed that their opponents did not see, nor had any knowledge of, these displays. They decided whether or not to compete for money against varying-ability opponents, by selecting an aggressive or non-aggressive display. We found that weaker opponents actually direct more competition towards the irrelevant aggressive display. Because displays were strategically irrelevant, we refer to this competitive behaviour as "non-instrumental". Such a 'goading effect' may reflect reactive aggression and/or a social norm against aggressive exploitation.

1. Introduction

Aggressive display behaviours are ubiquitous in the animal kingdom (Briffa and Hardy, 2013; Beaver, 2011). Human aggressive displays are alike across all known human societies and include a specific facial – particularly eyebrow – expression (Ekman and Friesen, 2003), vocal frequency and volume (van Staaden et al., 2011). According to classical ethology (Lorenz, 1966), aggressive displays evolved as a substitute for outright combat over contested resources: They provide valid information about social dominance - competitive ability and 'intent' - and therefore help discourage weaker contestants from futile competition. According to this evolutionary rationale – which has received support from evolutionary game theory (Enquist, 1985; Skyrms, 2009) – aggressive displays may reflexively trigger a submissive program, reminiscent of the 'involuntary defeat strategy' (Gilbert, 2000; Weisfeld and Wendorf, 2000). We wanted to know whether completely uninformative aggressive displays indeed discourage weaker opponents from competition. Alter-

natively, we wondered if aggressive displays were potent enough to trigger unconditional submission (Gilbert, 2000; Weisfeld and Wendorf, 2000) or, at the other extreme, unconditional 'defensive attack' Blanchard et al. (1980).

Players visited a behavioural lab in groups and participated in a variant of the 'hawk dove' conflict as follows. They were anonymously paired and simultaneously chose whether or not to compete for money in an 'intelligence contest'. If both players competed, the winner took 10$ and their opponent lost 10$. If neither player competed, a coin toss determined who took 10$, otherwise the player who chose to compete took it, see Fig 3. The principle of the game is that while each player prefers not to submit, competition can incur losses.

Players always knew their winning probability ω against their current opponent: it was displayed numerically on their screen on each trial. They chose to the COMPETE or NOT COMPETE response option with the mouse cursor. These response options were always labeled with boxes. For the 50% of players in the *no-display group*, response options were labeled with gray boxes. The remaining *display group* players concern us here. In this group one aggressive and one neutral face labeled the two response options respectively: the option COMPETE was either labeled by an aggressive display or a non-aggressive display, see Fig 2. Thus, on a random half of trials, COMPETE was labeled with an aggressive display and NOT COMPETE was labeled with a non-aggressive display: vice versa for the other half of trials. In this way, opponents varied in winning probability (relevant) and they were labeled with an aggressive or non-aggressive display (irrelevant). We wanted to know whether subjects directed more competition towards weaker opponents – 'instrumental competition' based on winning probability – and whether weaker opponents showed non-instrumental competition, i.e. directed less competition towards irrelevant aggressive dis-

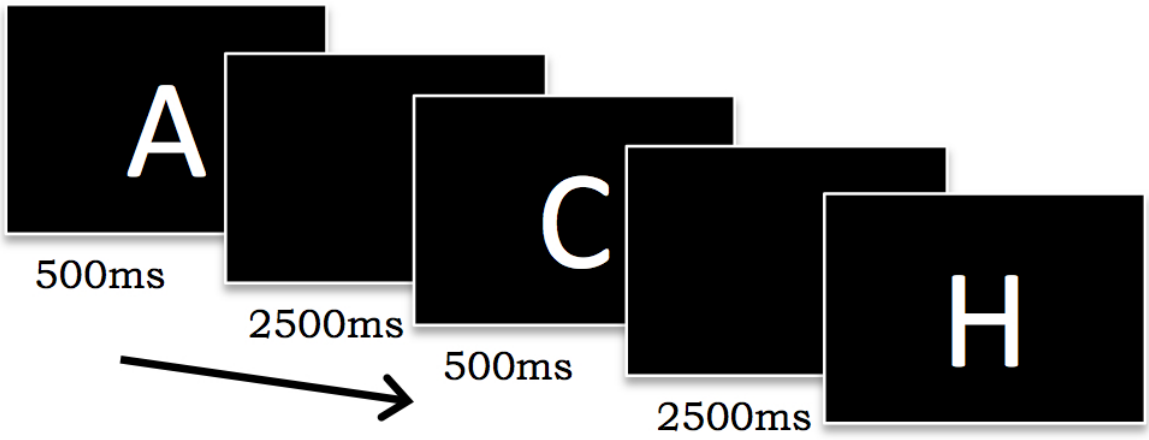

Figure 1: Fig 1. Screen during the N-Back working memory task.

plays.

2. Methods

2.1. Experiment 1

2.1.1. Subjects

The experiment was conducted in a computer laboratory at Zurich University. A total of 102 subjects (18-30 years old, 40 women) were tested in four sessions containing even-numbered groups of 20 to 34 subjects. The study was approved by the local ethics committee. Subjects were not deceived in any part of this study. Subjects payments depended on their real performance and choices in the task.

2.1.2. Procedure

Subjects were welcomed into a reception hall. Having been identified and instructed of the ground rules (see below), they were conveyed *en masse* into a separate behavioural lab, where they were each randomly assigned to an isolated computer booth. Subjects could only see their own screen and communication was prohibited. With no mention of the upcoming interactive

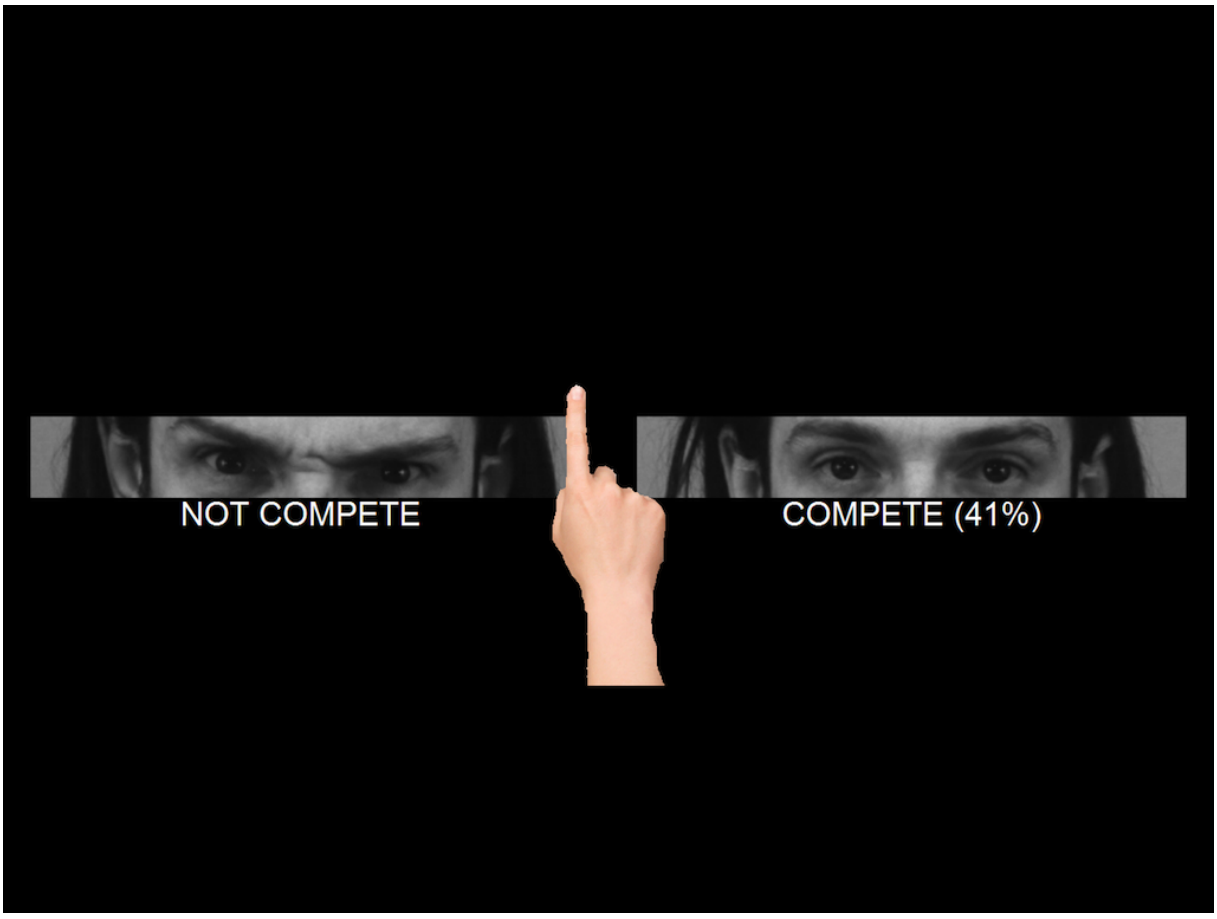

Figure 2: Fig 2. Screenshot for display group. This depicts a 'non-aggressive display' trial, in which a weaker opponent – winning probability 41% – may choose to compete by approaching a non-aggressive display. Subjects used the mouse cursor (hand) to click on either the left or the right display.

	COMPETE	NOT COMPETE
COMPETE	$10(1-\omega)-\omega 10$ $10\omega-10(1-\omega)$	0 10
NOT COMPETE	10 0	5 5

Figure 3: Fig 3. Expected payoffs in the game design. The row player chooses a row, similarly for the column player. These choices jointly determine which cell of this 2 x 2 table subjects end up in. The row players expected payoff in that end state is given (bottom) left of each cell: The column player's is given (top) right. ω is the row player's winning probability against the column player. $1-\omega$ is therefore the column player's winning probability.

task, they were first given instructions on how to perform the non-interactive, N-back test. For the full instructions given to subjects on the N-back, see Supplementary Fig 7.

In the *N-Back task*, subjects saw a sequence of stimuli and responded whenever a stimulus was the same as N steps before (Owen et al., 2005; Jaeggi et al., 2010). Specifically, subjects were shown a random sequence drawn from a pool of eight different letters (A-H). Each letter was shown for 500ms with an interval of 2500ms between two letters. Subjects were required to report 'target' letters by pressing the space key in the 3000ms interval before the next letter was shown. If the letter shown was not a target, a non-response was required. After a 'comprehension test', in which every subject was required to successfully complete a 1-back version, subjects completed three blocks with increasing difficulty, in which N was 2,3 and 4. Each block contained 20 stimuli and 4 targets, resembling the parameters of (Owen et al., 2005; Jaeggi et al., 2010). The whole N back task lasted approximately 15 minutes.

When all subjects had completed the N-back test, they received instructions on the interactive task. Most of the instructions were common to all subjects, see supplementary Fig 8. Briefly, subjects were given 25$ each and instructed that they would be asked to decide whether to compete with a string of opponents. They were instructed that one of these trials would be randomly selected at the end for real financial pay. Both players choices on that random trial would determine payment: if both chose to COMPETE, the contest would be decided by the pair-specific winning probability on that trial i.e. based on relative N-back performance. Subjects were then unknowingly randomized into two groups: the *no-display group* and *display group*. The *no-display group* was not instructed about, nor saw, any faces. The *display group* was instructed:

> "You will see images of faces on the screen. Your choices should not be influenced by these faces. These faces are completely irrelevant to your earnings. They do not depict your partner. None of your partners will see these faces. Your partners do not know that you see faces."

Thus, to prevent the display group from strategizing about displays – e.g. anticipating their op-

ponent's reaction to displays – they were always paired with no-display group players and were instructed that their opponents had no knowledge of the displays, nor saw displays themselves. This critical feature of our design is discussed below. To prevent learning or reputation effects, no player saw their opponent's choices until the end.

Display group subjects saw faces in every trial of the interactive task, see Fig 2. Faces depicted only males and were cropped so that only the eyes region were visible (Lundqvist et al., 1998). On each decision screen, the aggressive and non-aggressive displays depicted one specific individual's photograph. In this way they differed only in the aggressive display *per se*.

The game structure was the same for both the display and no-display groups, see Fig 3. In each trial, subjects were randomly and anonymously paired with a member of the other group. On any one trial, both subjects independently decided if they wanted to COMPETE or NOT COMPETE. Both subjects were always given their conditional winning probability (i.e. their probability of winning if both chose to COMPETE). The winning probability was always depicted underneath the COMPETE option, see Fig 2. We calculated the winning probability for each trial from the two players' relative performance on the preceding N-back working memory task. This represented the chance that each player would be correct if a random trial from their N-back performance was compared (see Supplementary material 5.1 for derivation). If player 1 had superior N-back performance to player 2, they would see a higher winning probability ($> 50\%$): player 2 would see the complementary probability, i.e. $< 50\%$. In total, subjects encountered each opponent twice, once with the aggressive/non-aggressive display labeling COMPETE. The individual photographed in these two trials was identical (only the labeling changed) and specific to those two trials, i.e. each real human opponent was labeled with photographs of exactly one person.

Subjects made their choices without any time limit. A new trial was only initiated after every subject had made their choice. After completing all their choices, subjects received feedback about their financial outcome on their computer screen. There then received and signed a receipt for this amount at their booth, before proceeding individually to a departure foyer where they were paid and dismissed.

Communication between computers was achieved via a basic server-client setup developed

in JAVA and MATLAB. For displaying stimuli, we used the additional Cogent Toolbox from the Laboratory of Neurobiology at University College London[1].

2.1.3. Statistical analysis

We used mixed effects regression to assess the effect of *winning probability* and/or *aggressive displays* on competing in the display group. This multilevel regression framework estimates the effect of *winning probability* and *aggressive displays* within-subject, before pooling this information to infer the average population effect. More specifically, we regressed subject's choice to compete on $\beta_0 + \beta_1 a + \beta_2 \omega + \beta_3 a\omega$, where a indicates trials in which the aggressive display labeled COMPETE and ω is the subject's winning probability, as detailed next.

Player i's winning probability on trial j, denoted ω_{ij}, depended on their N-back performance, relative to their opponent on that trial. In the Supplementary material, we discuss other role of other game theoretic factors in 'instrumental competition'. An indicator variable a_{ij} indexed trials in which the aggressive display labeled COMPETE. This permitted us to identify 'non-instrumental' behaviour triggered by the aggressive display, i.e. whether COMPETE was less likely when labeled with the aggressive display (see below). The interaction term $\omega_{ij}a_{ij}$ served to identify whether such non-instrumental behaviour varied between weaker and stronger opponents, denoted by ω. The multilevel logistic regression framework accommodates repeated-measures i.e. correlated choices within-player (Gelman, 2007): Each player had 4 parameters, denoted by the 4-vector $\beta_i = (\beta_{0i}, \beta_{1i}, \beta_{2i}, \beta_{3i})$. Assuming the β_i are drawn from a Gaussian population distribution, this gives Equation 1

[1]Cogent Graphics developed by John Romaya at the LON at the Wellcome Department of Imaging Neuroscience. http://www.vislab.ucl.ac.uk/cogent.php

$$
\begin{aligned}
P(y_{ij} = 1) &= logit^{-1}(\eta_{ij}) \qquad (1)\\
\eta_{ij} &= \beta_{0i} + \beta_{1i}a_{ij} + \beta_{2i}\omega_{ij} + \beta_{3i}a_{ij}\omega_{ij}\\
\beta_i &\sim N(\Theta, \Sigma)
\end{aligned}
$$

where y_{ij} is the player i's choice and equals 1 if and only if they choose to COMPETE in contest j, $logit^{-1}$ is the inverse logistic function and $\beta_i \sim N(\Theta, \Sigma)$ means that the random effects β_i are distributed according to a Gaussian probability distribution with mean $\Theta = (\beta_0, \beta_1, \beta_2, \beta_3)^T$ and 4×4 unrestricted covariance Σ. The 'group-level' parameters $\Theta = (\beta_0, \beta_1, \beta_2, \beta_3)^T$ quantify winning probability and display effects *on average in the population* and are therefore the object of statistical inference. This gives a simple random-intercept, random-slope model (Gelman, 2007).

For interpretation, we centered the winning probability ω_{ij} on the indifference point – the point at which subjects choose COMPETE and NOT COMPETE with equal probability – before multiplying it by 100, to give a percentage. Centering permits us to interpret β_1 as non-instrumental behaviour at indifference, i.e. the additional tendency for to direct hawkish choices towards aggressive, as opposed to non-aggressive, displays. β_3 quantifies whether this non-instrumental behaviour varies with winning probability. For reference, β_2 gives aggression directed towards non-aggressive displays, as a function of winning probability. We estimated this model using ReML in the statistical environment R.

For numerical stability in the estimation of Equation 1 we were obliged to exclude extreme winning probabilities ω. Specifically we excluded the top and bottom 5% of ω. This did not result in the exclusion of any subject.

2.2. Experiment 2

2.2.1. Rating experiment

We wanted to measure subjective evaluations to the aggressive and non-aggressive displays. Thirty additional subjects (17 women) therefore participated in another session. For each of the

aggressive and non-aggressive displays we used above, subjects rated how *interesting*, *pleasant* and *annoying* the stimulus was. Stimuli were displayed on a computer monitor and subjects used a mouse to rate each picture from *not at all* to *very* on a continuous scale, see Fig 5. The stimuli presentation order was randomized between subjects. Each choice was self-paced and subjects were paid a fixed 10$ for this task.

3. Results

3.1. Experiment 1

Choices: Subjects competed on 63% of all trials. On average over all trials and conditions, subjects competed above chance when their winning probability was above 45% and below chance when it was below 45%, i.e. they became *indifferent* between COMPETE and NOT COMPETE at 45%. We identified this indifference point by fitting all choices to a logistic function of winning probability before identifying the winning probability which implied 50% competing probability. As discussed above, we re-centered ω about this indifference point before proceeding with statistical inference.

Statistical inference derives from Equation 1. Winning probability ω significantly predicted competing probability ($p = 6.7 \times 10^{-12}, \hat{\beta_2} = .87, n = 51$). Subjects expressed significant non-instrumental competition at the indifference point ($p = 0.002, \hat{\beta_1} = 1.1, n = 51$), i.e. they directed more competitive behaviour towards the aggressive display than the non-aggressive display. This non-instrumental behaviour declined with winning probability, as indicated by a significant $c\omega$ interaction ($p = 2.3 \times 10^{-6}, \hat{\beta_3} = -0.23, n = 51$). Fig 1 visualizes this relationship between winning probability and competing probability, separately for aggressive display versus non-aggressive display trials. It shows that subjects were competitive against evenly-matched opponents, regardless of the display: when $\omega = 1/2$, competing probability was close to 100% on average. It further shows that, at the indifference winning probability (45%), subjects expressed non-instrumental competition: more competition towards aggressive than non-aggressive displays.

Reaction speeds: On average subjects took 3.14 seconds to decide. We asked if winning prob-

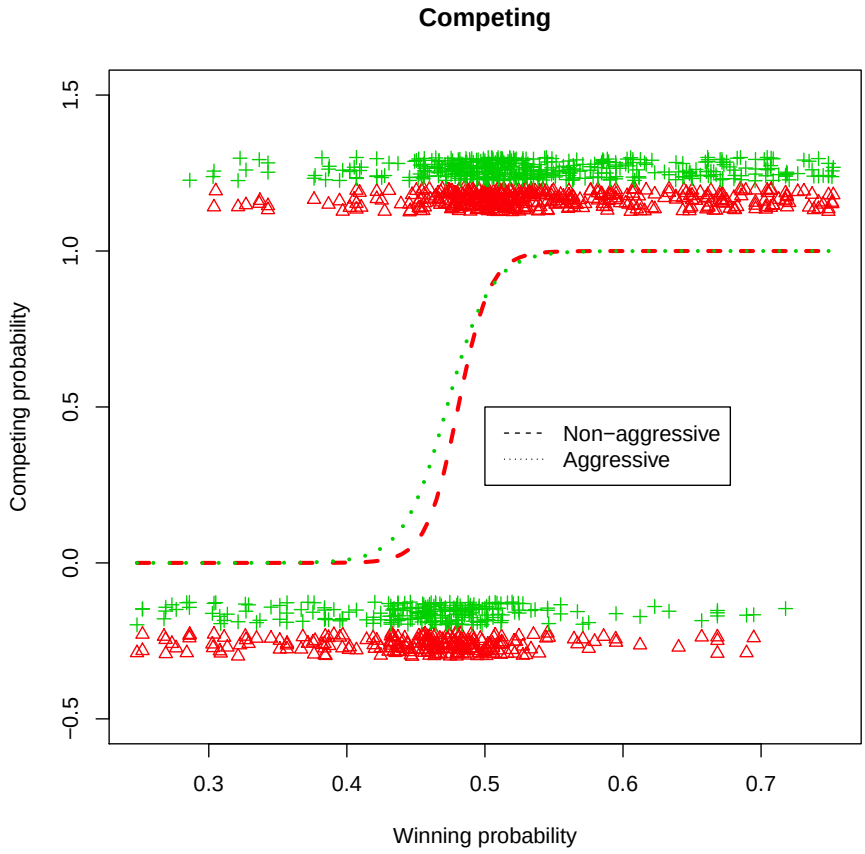

Figure 4: Fig 4. Competing as a function of winning probability and aggressive display. COMPETE choices are indicated by jittered points above 1 on the vertical axis. NOT COMPETE choices are indicated beneath 0. Green crosses indicate "aggressive display" trials, in which players were required to approach an aggressive display in order to COMPETE. Red triangles indicate in non-aggressive display trials, in they were required to approach a non-aggressive display in order to COMPETE. The dotted green line gives competing probability as a function of winning probability for aggressive display trials: where competing required approaching an aggressive display. The dashed red line gives the corresponding relationship when competing required approaching a non-aggressive display.

ability and/or display affected log reaction time. Specifically, we re-estimated Equation 1, having replaced line 1 with $p(log(RT_{ij})) = \eta_{ij}$ and introduced independent, zero-mean Gaussian errors ϵ_{ij} to line 2 giving $\eta_{ij} = \beta_{0i} + \beta_{1i}a_{ij} + \beta_{2i}\omega_{ij} + \beta_{3i}a_{ij}\omega_{ij} + \epsilon_{ij}$. This analysis revealed that reaction times somewhat quickened with both ω and a, but neither effect was statistically significant.

3.2. Experiment 2

Subjects face ratings were translated onto a scale from 0 to 1 for each picture. Using a paired t-test, we found no evidence for the hypothesis that the non-aggressive and the aggressive display were differently 'interesting', a proxy for 'salience'. Using a paired t-test, the aggressive display was significantly less 'pleasant' ($n = 30$, $p< 0.001$) and significantly more 'annoying' than the non-aggressive display ($n = 30$, $p< 0.001$). The resulting ratings are visualized in supplementary Fig 6.

4. Discussion

Players in our contest were 'instrumentally competitive': they competed more against weaker opponents. Weaker opponents in turn expressed 'non-instrumental competition': they directed more competitive behaviour towards irrelevant aggressive displays. Importantly, this result is based on our *within-subject* design and statistical analysis: non-instrumental competition cannot reflect between-subject strategic or socio-emotional variables, such as intelligence or personality.

Does this non-instrumental competition reflect a simple approach heuristic – 'approach aggressive display' – or a simple stimulus-response 'congruence' heuristic? This seems unlikely. First, by our definition, non-instrumental competition cannot reflect a general tendency to 'approach aggressive displays', either because they are salient (Hoffman, 1978; Jarvenpaa, 2011) or because they provoke directed defensive reflexes (Fanselow, 1992, 1994). By our definition, non-instrumental competition requires this approach behaviour to be competitive. Thus, due to counterbalancing, any subject who unconditionally approached the aggressive display would compete on exactly 50% of trials. They could not express *augmented* competing on aggressive display trials, i.e. non-instrumental competition. It is important to exclude this possibility because aggressive

displays can trigger involuntary autonomic responses (Eastwood and Smilek, 2005; Dimberg et al., 2000) as well as behavioural reflexes such as orienting and freezing (Dimberg et al., 2000). Second, our non-instrumental competition is unlikely to reflect subjects' preference for response-congruent stimuli *per se*, either due to 'priming effects' (Anderson and Bushman, 2001) or 'stimulus response compatibility' (Kornblum et al., 1990). On our aggressive display trials *both* options, COMPETE and DON'T COMPETE, were labeled with congruent displays. In other words, COMPETE was congruent with the aggressive display and NOT COMPETE was congruent with the non-aggressive display: congruence alone did not favor one response option. Further speaking against these approach and stimulus-response congruence heuristics, we observed that non-instrumental competition was somewhat context-specific, targeted more towards stronger opponents.

Does non-instrumental competition reflect some kind of strategizing? In nonhuman species, the behavioural response to aggressive displays is often strength- or rank-dependent (Chapais et al., 1994). Such context-sensitivity may simply reflect valuable strategic information within the displays themselves, i.e. displays convey the opponents fighting probability, winning probability or strategy (Blair, 2003). To exclude this possibility a critical feature of our design is that displays carried no information about the opponents strength or strategy: displays were randomly presented and subjects knew their opponents did not see faces and that their opponents did not know that they saw faces. Had we chosen not to instruct subjects at all, or to instruct subjects that displays depicted their opponents, any behavioural effect of displays may plausibly have reflected their attempt to second-guess the effect of displays on their opponents strategy or to second-guess their opponent's signaling motives, respectively. While it always remains possible that our instruction may itself have influenced participants' strategy, it is difficult for us to imagine any such "experimenter demand effect" in this experiment so we view this possibility sceptically.

Why might subjects express non-instrumental competition in this way? One possibility relates to the dichotomy between 'impulsive', 'reactive' 'affective' aggression and 'instrumental' or 'predatory' aggression (Weinshenker and Siegel, 2002; Bushman and Anderson, 2001). In contrast to 'instrumental aggression', where aggression is simply an 'instrument' used to acquire the contested resource, 'impulsive' aggression is a 'relatively automatic' response e.g. to threats (Beaver,

2011). While this distinction is notoriously vague, it begins to acknowledge different motives for aggression. It is relevant for us that aversive events – frustrations or unpleasant stimuli – can increase aggressive motivation by producing negative affect (Berkowitz, 1993; Anderson and Bushman, 2002). Our subjects viewed aggressive displays as unpleasant and annoying. Furthermore, low winning probability (or social status) may have been frustrating. Such an account however, would seem to predict an unconditional, main effect of the aggressive display – meaning that the aggressive display should increase "compete" playing throughout the zone of probabilistic playing. That we find an interaction –that the aggressive display disproportionately affects 'weaker' players – appears to require an account goes beyond unsophisticated, unconditional "reflexive" cognitive architecture. Alternatively, it reinforces the point that many reflexes, even spinal reflexes, are highly context-sensitive (Shemmell et al., 2010). Indeed in the context of primate social behaviour, defensive responses are highly context-sensitive, being dependent on an individual's relative social rank. Yet it remains difficult to conclusively demonstrate that our observations indeed reflects context-sensitive 'reflexes' to social stimuli, as previously hypothesized (Gilbert, 2000; Weisfeld and Wendorf, 2000). Thus many questions remain for future work, to establish whether this conjunction of aversive conditions generated motivation to attack the display and/or to challenge the opponent. In sum, a more clear and explicit explanation is needed for what the eye manipulation is doing to the proximate psychology and whether pressing the compete button with aggressive display is actually measuring defensive aggression.

Craig A Anderson and Brad J Bushman. Effects of violent video games on aggressive behavior, aggressive cognition, aggressive affect, physiological arousal, and prosocial behavior: A meta-analytic review of the scientific literature. *Psychological science*, 12(5):353–359, 2001.

Craig A Anderson and Brad J Bushman. Human aggression. *Annual review of psychology*, 53(1): 27–51, 2002.

Kevin M Beaver. *Aggression, advances in genetics, volume 75. Edited by Robert Huber, Danika L. Bannasch, and Patricia Brennan, x+ 296 pp. Boston, MA: Academic Press (Elsevier).* Wiley Online Library, 2011.

Leonard Berkowitz. *Aggression: Its causes, consequences, and control.* Mcgraw-Hill Book Company, 1993.

R. J. R. Blair. Facial expressions, their communicatory functions and neuro-cognitive substrates. *Philosophical Transactions: Biological Sciences*, 358(1431):pp. 561–572, 2003. ISSN 09628436.

R.J. Blanchard, C.F. Kleinschmidt, C. Fukunaga-Stinson, and D.C. Blanchard. Defensive attack behavior in male and female rats. *Learning & Behavior*, 8(1):177–183, 1980.

Mark Briffa and Ian CW Hardy. Introduction to animal contests. *Animal contests*, pages 1–4, 2013.

Brad J Bushman and Craig A Anderson. Is it time to pull the plug on hostile versus instrumental aggression dichotomy? *Psychological review*, 108(1):273, 2001.

Bernard Chapais, Jean Prud'Homme, and Shona Teijeiro. Dominance competition among siblings in japanese macaques: constraints on nepotism. *Animal behaviour*, 48(6):1335–1347, 1994.

U. Dimberg, M. Thunberg, and K. Elmehed. Unconscious facial reactions to emotional facial expressions. *Psychological Science*, 11(1):86–89, 2000.

John D. Eastwood and Daniel Smilek. Functional consequences of perceiving facial expressions of emotion without awareness. *Consciousness and Cognition*, 14(3):565 – 584, 2005. ISSN 1053-8100.

Paul Ekman and Wallace V Friesen. *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk, 2003.

Magnus Enquist. Communication during aggressive interactions with particular reference to variation in choice of behaviour. *Animal Behaviour*, 33(4):1152–1161, 1985.

M.S. Fanselow. The midbrain periaqueductal gray as a coordinator of action in response to fear and anxiety. *The midbrain periaqueductal gray matter*, pages 151–173, 1992.

M.S. Fanselow. Neural organization of the defensive behavior system responsible for fear. *Psychonomic Bulletin & Review*, 1(4):429–438, 1994.

N. Feltovich. The effect of subtracting a constant from all payoffs in a hawk-dove game: experimental evidence of loss aversion in strategic behavior. *Southern Economic Journal*, 77(4):814–826, 2011.

S. Gachter, E. Johnson, and A. Herrmann. Individual-level loss aversion in riskless and risky choices. 2007.

Andrew Gelman. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2007.

P. Gilbert. Varieties of submissive behavior as forms of social defense: Their evolution and role in depression. *Subordination and defeat: An evolutionary approach to mood disorders and their therapy*, pages 3–45, 2000.

J.E. Hoffman. Search through a sequentially presented visual display. *Attention, Perception, & Psychophysics*, 23(1):1–11, 1978.

S.M. Jaeggi, B. Studer-Luethi, M. Buschkuehl, Y.F. Su, J. Jonides, and W.J. Perrig. The relationship between n-back performance and matrix reasoning-implications for training and transfer. *Intelligence*, 38(6):625–635, 2010.

S.L. Jarvenpaa. Graphic displays in decision making – the visual salience effect. *Journal of Behavioral Decision Making*, 3(4):247–262, 2011.

D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, pages 263–291, 1979.

Sylvan Kornblum, Thierry Hasbroucq, and Allen Osman. Dimensional overlap: cognitive basis for stimulus-response compatibility–a model and taxonomy. *Psychological review*, 97(2):253, 1990.

Konrad Lorenz. *On Aggression. translated by marjorie kerr wilson*. Harcourt, Brace & World, 1966.

Daniel Lundqvist, Anders Flyk, and Arne Öhmann. *The Karolinska Directed Emotional Faces - KDEF*. CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institute, 1998.

J. Maynard Smith. The theory of games and the evolution of animal conflicts. *Journal of theoretical biology*, 47(1):209–221, 1974.

A.M. Owen, K.M. McMillan, A.R. Laird, and E. Bullmore. N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human brain mapping*, 25(1):46–59, 2005.

Jonathan Shemmell, Matthew A Krutky, and Eric J Perreault. Stretch sensitive reflexes as an adaptive mechanism for maintaining limb stability. *Clinical Neurophysiology*, 121(10):1680–1689, 2010.

Brian Skyrms. Evolution of signalling systems with multiple senders and receivers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1518):771–779, 2009.

J.M. Smith. *Evolution and the Theory of Games*. Cambridge university press, 1982.

J.M. Smith and GR Price. The logic of animal conflict. *Nature*, 246:15, 1973.

R. Thaler. Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization*, 1(1):39–60, 1980.

Amos Tversky and Daniel Kahneman. Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, 106(4):pp. 1039–1061, 1991. ISSN 00335533.

330 Moira J van Staaden, William A Searcy, and Roger T Hanlon. Signaling aggression. *Advances in*
331 *genetics*, 75:23–49, 2011.

332 Naomi J Weinshenker and Allan Siegel. Bimodal classification of aggression: affective defense and
333 predatory attack. *Aggression and Violent Behavior*, 7(3):237–250, 2002.

334 G. Weisfeld and C. Wendorf. The involuntary defeat strategy and discrete emotions theory. *Sub-*
335 *ordination and defeat: An evolutionary approach to mood disorders and their therapy*, pages 121–145,
336 2000.

5. Supplementary Material

5.1. Calculating the winning probability ω

A players N-back performance might be given by

$$perf = \frac{\text{True positives}}{\text{Targets}} + \frac{\text{True negatives}}{\text{Non-Targets}} \tag{2}$$

Assuming half of all trials contain a target, this gives the chance of being correct in a random N-back trial, $P(correct)$. More generally, $P(correct)$ is given by reweighting the two terms in $perf$ by proportion of targets/non-targets respectively. We wanted our winning probability ω to reflect *relative* N-back ability: namely the chance of one player winning in a randomly selected N-back trial. Thus, if neither or both players were correct on that trial, another trial is picked until there is a winner. Thus, ω_{12}, player 1's winning probability against player 2, is the chance of player 1 being correct, conditional on having a distinct winner (no tie), i.e.

$$\omega_{12} = p(player\,1\,correct|no\,tie)$$

Let $P(correct)$ of player 1 and 2 be $P1$ and $P2$ respectively. Then applying Bayes' Theorem gives

$$\omega_{12} = \frac{p(no\,tie|player\,1\,correct)P1}{p(no\,tie|player\,1\,correct)P1 + p(no\,tie|player\,1\,not\,correct)(1-P1)}$$

$p(no\,tie|player\,1\,correct)$ is the chance of having a winner, given that player 1 was correct. This can only be the case if player 2 was incorrect and consequently is given by $(1-P2)$. Similarly, $p(no\,tie|player\,1\,not\,correct)$ is the chance of having a winner, given that player 1 was incorrect. This can only be the case if player 2 was correct and is consequently given by $P2$. Substituting gives

$$\omega_{12} = \frac{(1-P2)P1}{(1-P2)P1 + (1-P1)P2} \tag{3}$$

where ω_{12} is player 1's winning probability against player 2.

In matrix notation, let $perf(p_1, p_2, ..., p_n)$ be a performance vector with p_1, p_2 .. p_n now representing individual subjects' N-back scores. Multiplying this vector with its complementary vector gives

$$perf^T * (1 - perf) = X_{n,n} = \begin{pmatrix} p_1(1-p_1) & p_1(1-p_2) & \cdots & p_1(1-p_n) \\ p_2(1-p_1) & p_2(1-p_2) & \cdots & p_2(1-p_n) \\ \vdots & \vdots & \ddots & \vdots \\ p_n(1-p_1) & p_n(1-p_2) & \cdots & p_n(1-p_n) \end{pmatrix}$$

If we divide the matrix X element-wise by $X + X^T$ we get

$$\frac{X}{X+X^T} = \begin{pmatrix} \frac{p_1(1-p_1)}{p_1(1-p_1)+p_1(1-p_1)} & \frac{p_1(1-p_2)}{p_1(1-p_2)+p_2(1-p_1)} & \cdots & \frac{p_1(1-p_n)}{p_1(1-p_n)+p_n(1-p_1)} \\ \frac{p_2(1-p_1)}{p_2(1-p_1)+p_1(1-p_2)} & \frac{p_2(1-p_2)}{p_2(1-p_2)+p_2(1-p_2)} & \cdots & \frac{p_2(1-p_n)}{p_2(1-p_n)+p_n(1-p_2)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{p_n(1-p_1)}{p_n(1-p_1)+p_1(1-p_n)} & \frac{p_n(1-p_2)}{p_n(1-p_2)+p_2(1-p_n)} & \cdots & \frac{p_n(1-p_n)}{p_n(1-p_n)+p_n(1-p_n)} \end{pmatrix} \tag{4}$$

$$= \begin{pmatrix} \omega_{11} & \omega_{12} & \cdots & \omega_{1n} \\ \omega_{21} & \omega_{22} & \cdots & \omega_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{n1} & \omega_{n2} & \cdots & \omega_{nn} \end{pmatrix} \tag{5}$$

which gives our winning probabilities for each pair of opponents, ij.

5.2. Other theoretical decision parameters

Our main analysis assumed that winning probability ω influences competitive behaviour. In

theory, other parameters might mediate 'instrumental competition'. For example, decision-theoretically optimal players with first-order beliefs will hold an expectation π_{ij} about whether their opponent will compete in this trial, and compete themselves only when the expected utility of competing is higher than not competing. Alternatively, game-theoretically optimal players - who view their opponents as rational - should agree on the Nash equilibrium in each trial. To derive trial-by-trial predictive quantities from either framework requires additional assumptions which we briefly discuss below.

Decision theory: Player i's best strategy in our task depends on what his opponent will do. Faced with a hawk, dove guarantees 0 gain/loss, while hawk will win with probability ω. Faced with a dove, hawk guarantees a gain. Either way, player i must simply choose the highest value response. But subjects cannot know their opponents strategy for sure. In our task, uncertainty about winning (from ω) is compounded by uncertainty about what one's opponent will do. If players expect their opponent to fight with probability π_{ij}, then under the Reduction of Compound Lotteries axiom of expected utility, this compound lottery reduces to a simple lottery. Sadly while the influence of π_{ij} on the expected utility can be quantified in this way, π_{ij} itself cannot easily be measured or inferred without strong assumptions.

Game theory: The hawk-dove game was first introduced as a model of animal conflict (Smith and Price, 1973) in which two animals compete over a resource V. If an aggressive hawk meets a submissive dove, the hawk takes the uncontested resource. A hawk will always fight another hawk until a winner is decided, at which point the loser pays cost C. When two doves meet they share the resource equally or – if the resource is indivisible – they display for a random period of time. (Smith, 1982) The dove that displays the longest gets the resource. This so called *war of attrition* (Maynard Smith, 1974) guarantees an optimal strategy where the winner is decided by chance. The expected payoffs of that game are given in the payoff Fig 10.

From this model one can derive the proportion of doves and hawks that evolve in a *population*. Alternatively, one can derive the probability with which two strategic competitors will choose between two options *hawk* and *dove*. Assuming loss aversion in our game (see below), $V < C$, the hawk-dove game takes the form of an anti-coordination game with two pure equilibria at

(*dove*,*hawk*) and (*hawk*,*dove*). An additional mixed equilibrium exists when we allow players to play a mixed strategy where they sometimes play *hawk* and sometimes play *dove*.
Let

$$\begin{aligned} p &= \text{probability of opponent playing } \textit{hawk} \\ (1-p) &= \text{probability of opponent playing } \textit{dove} \\ E(H,D) &= \text{expected payoff of playing } \textit{hawk} \text{ against } \textit{dove}. \\ &\quad \text{(similar notion for other strategy pairs)} \\ \pi(S) &= \text{total expected payoff of strategy } S \end{aligned}$$

In a Nash equilibrium no player has an incentive to deviate from his strategy. Therefore the probability of p is implicitly given by the equation that sets the other player indifferent between the two options:

$$\begin{aligned} \pi(H) &= \pi(D) \\ p * E(H,H) + (1-p) * E(H,D) &= p * E(D,H) + (1-p) * E(D,D) \end{aligned} \tag{6}$$

Inserting the payoffs of Fig 10 gives

$$\begin{aligned} p * \frac{1}{2}(V-C) + (1-p)V &= p * 0 + (1-p)\frac{V}{2} \\ p &= \frac{V}{C} \end{aligned} \tag{7}$$

Which of these equilibria are 'evolutionary stable strategy' (ESS): i.e. robust to invasion by another strategy. The answer depends on whether players can distinguish between their roles, meaning if they know that they are a row player or a column player. If this distinction can be made, an *uncorrelated asymmetry* exists. This is necessary for a pure ESS in an anti-coordination game (Maynard Smith, 1974). Thus contestants can establish a convention such that one always plays *hawk*, the other *dove*. Thus the two equilibria in pure strategies are evolutionary stable and

the mixed equilibrium is unstable. If there is no such asymmetry, then the mixed strategy is the only ESS. We return to this distinction below.

To better model our game, we next extend the hawk-dove game above, explicitly allowing for different winning probabilities and loss aversion (Kahneman and Tversky, 1979; Thaler, 1980), in which "the loss of utility associated with giving up a valued good is greater than the utility gain associated with receiving it" (Tversky and Kahneman, 1991). Individual loss aversion in risky choices is on average 1.5 (Gachter et al., 2007). Since subjects receive a 25$ endowment in our game, this constant is added to each outcome and the expected payoff of (COMPETE,COMPETE) remains in the positive domain. (Feltovich, 2011) showed that loss aversion persists when all hawk-dove payoffs are moved into the positive domain. Consequently we introduce a loss aversion parameter $A > 1$ into the payoff matrix in Fig 11. Following the preceding discussion, this ensures that there is no dominant strategy when opponents are equally matched. Second, the winning probabilities for opponents i and j at (*hawk*,*hawk*) are given by $\omega_i + \omega_j = 1$ such that both are ≥ 0 and ≤ 1.

In addition to the two pure equilibria, a single mixed strategy equilibrium can be calculated for this game. However, as noted above, if ω_{ij} is known to both players and $\omega_{ij} \neq 1/2$, then the game is asymmetric and this mixed strategy is unstable. This mixed strategy, where no distinction is made between players, gives lower average payoff than a pure strategy based on assigning conventional roles to players (Maynard Smith, 1974). Therefore, the only possible ESS in the asymmetric game are

$$ESS = \begin{cases} (dove,hawk) & \text{for } \omega_1 \leq \frac{V}{V+AC} \\ (dove,hawk) \text{ or } (hawk,dove) & \text{for } \frac{V}{V+AC} < \omega_1 < \frac{AC}{V+AC} \\ (hawk,dove) & \text{for } \omega_1 \geq \frac{AC}{V+AC} \end{cases}$$

An intuitive convention in our game is for the contestant with higher winning probability to play *hawk*, and the other contestant to play *dove*. However, this would imply that subjects completed on average 50% of the time, which contradicts our observation.

6. Figure legends

Supplementary Fig 5. Screenshot during the face rating.

Supplementary Fig 6. Mean self-report ratings for aggressive and non-aggressive displays. This graph shows that subjects found aggressive displays significantly less pleasant and more annoying, but not significantly more interesting.

Supplementary Fig 7. Instructions for the non-interactive N-back test. In addition to verbal instructions, subjects read this information slide from their computer monitor.

Supplementary Fig 8. Instructions for the interactive 'Hawk-Dove' game. In addition to verbal instructions, subjects read this information slide from their computer monitor. Additional instructions were given to the display group, as indicated in the main text.

Supplementary Fig 9. Comprehension test for the interactive 'Hawk-Dove' game. Subjects completed this test before proceeding to the task proper.

Supplementary Fig 10. General payoff matrix of the hawk-dove game.

Supplementary Fig 11. Hawk-dove game with winning probability ω and loss aversion A.

7. Figures

Figure 5:

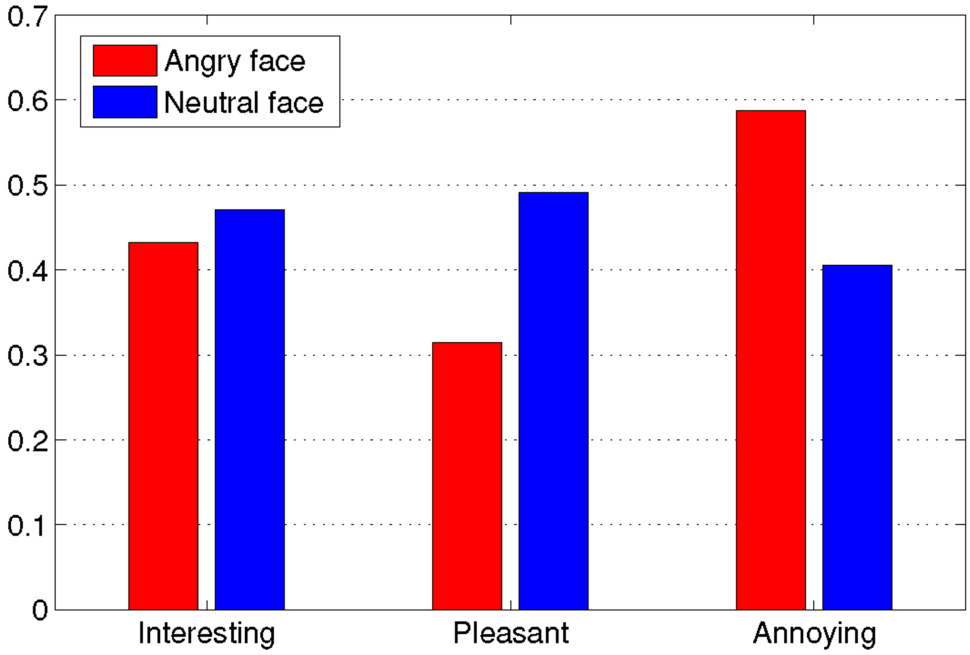
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Angry face
Neutral face
Interesting
Pleasant
Annoying

Figure 6:

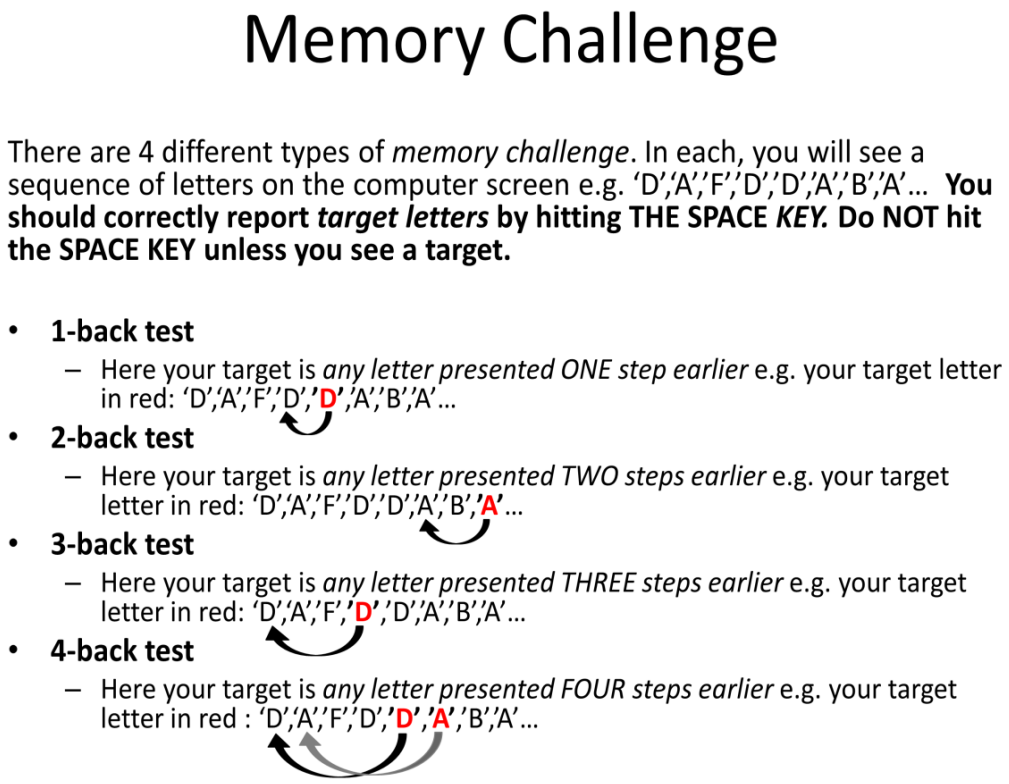
Memory Challenge
There are 4 different types of *memory challenge*. In each, you will see a sequence of letters on the computer screen e.g. 'D','A','F','D','D','D','A','B','A'… **You should correctly report *target letters* by hitting THE SPACE *KEY.* Do NOT hit the SPACE KEY unless you see a target.**
• **1-back test**
– Here your target is *any letter presented ONE step earlier* e.g. your target letter in red: 'D','A','F','D','D','A','B','A'…
• **2-back test**
– Here your target is *any letter presented TWO steps earlier* e.g. your target letter in red: 'D','A','F','D','D','A','B','A'…
• **3-back test**
– Here your target is *any letter presented THREE steps earlier* e.g. your target letter in red: 'D','A','F','D','D','A','B','A'…
• **4-back test**
– Here your target is *any letter presented FOUR steps earlier* e.g. your target letter in red : 'D','A','F','D','D','A','B','A'…

Figure 7:

Instructions

- You will be paired with ten other people in random sequence.
- You start the experiment with **25CHF**. You end with **15CHF**, **25CHF** or **35CHF**, depending on your choices. Please try to earn as much money as possible for yourself.
- To earn money, decide whether to **COMPETE** or **NOT COMPETE** against each person. If both of you compete, we will determine the winner by selecting one target from the previous n-back memory test
- You retain, loose or gain money as follows...
  - **If you both compete, the winner gains 10CHF, the loser loses 10CHF. To help you decide, your winning probability against the current partner is displayed on your screen. We calculated this from your relative performance in the memory test earlier.**
  - **If neither competes, the winner is determined by coin flip and gets 10CHF, the loser gets 0CHF.**
  - **If you compete and your partner does not compete, you get 10CHF with no contest; your partner gets 0CHF.**
  - **Conversely, if you do not compete and your partner competes, they get 10CHF uncontested; you get 0CHF.**
- At the end of the experiment, only 1 round will be selected at random for payment. You should therefore treat every round as if it were the only round that counts.

Figure 8:

Comprehension test

	What was each person's FINAL payoff?	
	Sally	Bob
1. Sally COMPETEs, bob does NOT COMPETE.	_____CHF	_____CHF
2. Sally does NOT COMPETE, bob COMPETEs.	_____CHF	_____CHF
3. Sally COMPETEs, bob COMPETEs → Sally wins the memory test.	_____CHF	_____CHF
4. Sally does NOT COMPETE, bob does NOT COMPETE → Bob wins the coin-flip.	_____CHF	_____CHF

1

Figure 9:

	COMPETE (*hawk*)	DON'T COMPETE (*dove*)
COMPETE (*hawk*)	$\frac{1}{2} * (V - C)$, $\frac{1}{2} * (V - C)$	V , 0
DON'T COMPETE (*dove*)	0 , V	$\frac{V}{2}$, $\frac{V}{2}$

Figure 10:

	COMPETE (*hawk*)	DON'T COMPETE (*dove*)
COMPETE (*hawk*)	$\omega_2 V - (1 - \omega_2)AC$ $\omega_1 V - (1 - \omega_1)AC$	0 V
DON'T COMPETE (*dove*)	V 0	$V/2$ $V/2$

Figure 11: