CeDEx

Centre for Decision Research and Experimental Economics

Discussion Paper Series

ISSN 1749-3293

CeDEx Discussion Paper No. 2008–17

Explaining Focal Points: Cognitive Hierarchy Theory *versus* Team Reasoning

Nicholas Bardsley, Judith Mehta, Chris Starmer and Robert Sugden

December 2008

The Centre for Decision Research and Experimental Economics was founded in 2000, and is based in the School of Economics at the University of Nottingham.

The focus for the Centre is research into individual and strategic decision-making using a combination of theoretical and experimental methods. On the theory side, members of the Centre investigate individual choice under uncertainty, cooperative and non-cooperative game theory, as well as theories of psychology, bounded rationality and evolutionary game theory. Members of the Centre have applied experimental methods in the fields of Public Economics, Individual Choice under Risk and Uncertainty, Strategic Interaction, and the performance of auctions, markets and other economic institutions. Much of the Centre's research involves collaborative projects with researchers from other departments in the UK and overseas.

Please visit http://www.nottingham.ac.uk/economics/cedex/ for more information about the Centre or contact

Karina Whitehead
Centre for Decision Research and Experimental Economics
School of Economics
University of Nottingham
University Park
Nottingham
NG7 2RD
Tel: +44 (0) 115 95 15620
Fax: +44 (0) 115 95 14159
karina.whitehead@nottingham.ac.uk

The full list of CeDEx Discussion Papers is available at

http://www.nottingham.ac.uk/economics/cedex/papers/index.html

Explaining Focal Points:
Cognitive Hierarchy Theory *versus* Team Reasoning[♦]

Nicholas Bardsley,[†] Judith Mehta,[*] Chris Starmer,[#] Robert Sugden[*]

[†]National Centre for Research Methods, University of Southampton
[*]School of Economics, University of East Anglia
[#]CeDEx, University of Nottingham

5 December 2008

[♦] The Amsterdam experiment was conducted while Bardsley was affiliated with CREED, Universiteit van Amsterdam; it was programmed by Jos Theelen and financed by EU TMR project 'ENDEAR', FMRX-CT98-0238. The Nottingham experiment, and Starmer and Sugden's subsequent work, were supported by the Leverhulme Trust; Jacinto Braga, Steve Humphrey and Henrik Orzen helped in running the experiment. We also thank Vincent Crawford, Peter Moffatt, Tassos Magdalinos, David Myatt, Thomas Schelling, Ignacio Zamarrón, two anonymous referees, and participants at the 2001 Barcelona meeting of the Economic Science Association for comments. The idea of using a 'guessing' treatment was first suggested by a student participant at a seminar at the University of Oxford, shortly after the publication of Mehta et al (1994); we regret that we do not know her name.

Explaining Focal Points:
Cognitive Hierarchy Theory *versus* Team Reasoning

Abstract

This paper reports experimental tests of two alternative explanations of how players use focal points to select equilibria in one-shot coordination games. Cognitive hierarchy theory explains coordination as the result of common beliefs about players' pre-reflective inclinations towards the relevant strategies; the theory of team reasoning explains it as the result of the players' using a non-standard form of reasoning. We report two experiments. One finds strong support for team reasoning; the other supports cognitive hierarchy theory. In the light of additional questionnaire evidence, we conclude that players' reasoning is sensitive to the decision context.

Keywords: salience, focal point, cognitive hierarchy, team reasoning

JEL codes: C72 (noncooperative games), C92 (laboratory, group behaviour)

It is well known that the players of one-shot coordination games are often successful to a degree that classical game theory cannot explain. In these games, particular Nash equilibria seem to constitute ‘focal points’ on which the players’ expectations converge. The existence of focal points was first demonstrated by Schelling (1960); his informal experiments have been replicated under controlled conditions (Mehta, Starmer and Sugden, 1994). Although the concept of a focal point has been routinely used in game theory for many years, there is still no generally-accepted explanation of how, in reality, real people manage to reach these equilibria. Two alternative lines of explanation have developed. One approach, first suggested by Lewis (1969), rests on assumptions about ‘primary salience’ – that is, players’ psychological propensities to play particular strategies by default, when there are no other reasons for choice. More recently, this approach has been formalised as *level-*n *theory* (Stahl and Wilson, 1995; Bacharach and Stahl, 2000) and, in a simplified form, as *cognitive hierarchy theory* (Camerer, Ho and Chong, 2004).[1] The other approach, arguably implicit in some parts of Schelling’s own analysis, assumes that each player chooses the decision rule which, if used by all players, would be optimal for each of them. This has since been formalised as the theory of *team reasoning* (Sugden, 1993, 1995; Bacharach, 1999, 2006). In this paper, we report two experiments designed to discriminate between these approaches to explaining coordination.

We begin by setting out the two approaches and showing that, for certain classes of coordination games, they make different predictions (section 1). We describe an experimental design which allows two tests of these predictions. The first test uses pure coordination games in which strategies are distinguished by labels, and compares the behaviour of subjects in three treatments: ‘pickers’, who choose between labels without any incentive to choose one rather than another, ‘guessers’, who guess how pickers have behaved, and ‘coordinators’, who try to coordinate with one another. Because the relevant cross-treatment comparisons can be made for *any* given set of labels, this test does not depend on prior assumptions about the salience of the labels used. This is a great advantage. There is a widespread perception among economists that when salience is culturally dependent, as it is in the most famous of Schelling’s games, it is resistant to decision-theoretic analysis. In our cross-treatment comparisons, the cultural determinants of salience are held constant, permitting direct tests of game-theoretic hypotheses about play in these games. The second test uses another type of coordination game discussed by Schelling, in which both players' payoffs are higher in some Nash equilibria than in others (section 2).

We implemented this design in two experiments, which differed only in apparently small details. Surprisingly, the results of one experiment seem to support the theory of team reasoning, while those of the other seem to support cognitive hierarchy theory (sections 3, 4 and 5). We investigated the reasons for this difference by using questionnaires to elicit perceptions of salience from members of the two subject pools (section 6). Reviewing our experimental results in the light of the questionnaire responses, we conclude that modes of reasoning similar to those modelled by each of the two theoretical approaches are at work; which of them is used is sensitive to subtle differences in the specifications of coordination tasks. The implication, we suggest, is that one should be pessimistic about finding any simple, unified theory of focal points. We suggest that this conclusion is consistent with Schelling's own analysis, which emphasises the diversity of the methods by which focal points are found (section 7).

We are conscious that this paper does not have a simple story line. It would have been easy for us to provide one, either by reporting each of the experiments separately, or by reporting only one of the two tests that the experiments were designed to conduct; but such a strategy would have given a false representation of what we know to be the case. As we shall explain, neither the differences between the results of the two experiments nor the differences between those of the two tests can be understood as revealing lack of control in the experimental design. We have carried out what we believe to be a well-controlled investigation of the two leading approaches to explaining a phenomenon that has puzzled game theorists for nearly half a century. Our findings may be disappointing to readers who are looking for a simple and general game-theoretic explanation of focal points, but – as Schelling warned from the outset – the whole idea of such an explanation may be no more than a mirage. If that is the case, it is important to know.

1 Theory

1.1 *The framework*

Throughout this paper, we are concerned only with one-shot coordination games.[2] Our aim is to understand how human players actually coordinate. For simplicity, we confine our attention to two-player games.

Our definition of a *coordination game* refers both to its normal form – which is how it is represented in classical game theory – and to the mechanism of 'labelling' which allows players to distinguish between strategies. Described by its normal form, a coordination game is a game for players 1 and 2. Player 1 chooses a strategy from the set $S_1 = \{s_{11}, \ldots, s_{1n}\}$ where $n \geq 2$; player 2 chooses from $S_2 = \{s_{21}, \ldots, s_{2n}\}$. Payoffs are defined in terms of a vector of strictly positive utility indices $U_1, ..., U_n$. If, for some j, the chosen strategies are s_{1j} and s_{2j} (that is, if both players choose strategies with the same index j), then each player receives the payoff U_j; otherwise, each receives zero. The case in which $U_1 = U_2 = ... = U_n$ is a *pure coordination game*; otherwise there is a *Hi-Lo game*.[3]

In the normal form of a pure coordination game, the n strategies of each player are completely symmetrical with one another; correspondingly, there are n symmetrical Pareto-efficient pure-strategy Nash equilibria (s_{1j}, s_{2j}). In classical game theory, each Nash equilibrium is treated as a candidate 'solution'; the players' problem is to 'select' one equilibrium from the set of candidates. However, if we consider only the normal form of the game, it is not clear what selection of a pure-strategy equilibrium can mean. If one equilibrium is to be singled out by the players, it must be distinguished from the others in some way that both of them can recognise; but if the n pure-strategy equilibria are completely symmetrical with one another, what distinguishes one from another? Indeed, some theorists claim that, in a world of ideal rationality, these equilibria *are* indistinguishable, and hence that the only rational solution to a pure coordination game is the mixed strategy equilibrium which assigns a probability of $1/n$ to each pure strategy (Harsanyi and Selten, 1988).

In theoretical analyses of focal points in pure coordination games, the problem of indistinguishability is usually overcome by making explicit assumptions about the labelling of strategies (Mehta et al, 1994; Bacharach and Stahl, 2000; Bacharach, 1993; Casajus, 2001). Following this approach, we make it part of the definition of a coordination game that there is a set $L = \{l_1, \ldots, l_n\}$ of distinct *labels*, common to both players; these may be words, numbers, pictures, rows or columns in a matrix, or anything else that players can recognise. In cases in which labels consist of strings of characters, we denote this by enclosing the relevant strings in the symbols « and »; thus the coordination game in which players name 'heads' or 'tails' can be denoted by $L = \{l_1, l_2\}$ with l_1 = «heads» and l_2 = «tails». Each player knows L, and registers her strategy choice by choosing a label from this set. Labels are tied to strategies so that, if player *i* chooses label l_j, she thereby chooses the strategy denoted s_{ij} in the normal form.

By using the concept of labelling, it is possible to talk meaningfully about equilibrium selection in pure coordination games. But the problem remains of explaining the remarkable success with which human players choose the same labels in these games.

Since we are trying to explain this success, it is useful to have an operational measure of it. We adapt a measure proposed by Mehta et al (1994). Consider a coordination game with the label set $L = \{l_1, ..., l_n\}$, and any set of N individuals, each of whom plays that game once with an anonymous co-player. For each label l_j, let m_j be the number of individuals who choose it. Then the *coordination index c* is given by:

(1) $c = \sum_j m_j (m_j - 1) / [N (N - 1)]$.

This index measures the probability that two distinct individuals, chosen at random without replacement from the set of N individuals, choose the same label. It takes the value 1 if all individuals choose the same label, and 0 if everyone chooses a different label. If labels are chosen at random, the expected value of the index is $1/n$. When making comparisons between games with different numbers of labels, it is clarifying to use the *normalised coordination index (NCI)* defined by $c^* = cn$. This can be interpreted as the ratio of c, the probability that two randomly-chosen individuals choose the same label, to $1/n$, the corresponding probability if labels are chosen at random. For example, Schelling (1960, pp. 54–58) reports an experiment in which 42 people were asked how they would choose in a pure coordination game with L = {«heads», «tails»}; 36 chose «heads». This implies $c = 0.75$, while random picking would imply the expected value $c = 0.5$. Thus $c^* = 1.50$. We will say that, among a population of players of a pure coordination game, the distribution of label choices is more *concentrated* (or, equivalently, less *dispersed*), the higher the value of c^*.

Clearly, any explanation of why NCIs in pure coordination games are consistently higher than 1 must take some account of the content of the labels: it must show how some labels are more attractive or choiceworthy than others. We now consider two alternative theoretical approaches, each of which allows choices to be influenced by labelling.

1.2 *Primary and secondary salience in pure coordination games: cognitive hierarchy theory*

The idea that focal points can be explained in terms of 'primary' and 'secondary' salience was first proposed by Lewis (1969, pp. 24-36).[4] It is now possible to formulate this hypothesis more rigorously in terms of cognitive hierarchy theory. We apply this theoretical approach to the case of a two-player coordination game. In such a game, each player chooses from the same set of labels L. A player's behaviour can be represented by a probability distribution over these labels, typically denoted by $\boldsymbol{p} = (p_1, \dots , p_n)$. (As we shall not need to refer to specific players, we dispense with the indices '1' and '2' which identify the players.)

The theory postulates a hierarchy of *cognitive levels* 0, 1, ... Each player has a specific cognitive level, representing the degree to which he can reason about other players. Players are uncertain about the cognitive levels of their opponents. For each level k, the relative frequency of level k players in the population of potential players of the game is q_k; it is required that $q_0 > 0$.

Level 0 reasoners do not use game-theoretic reasoning, but simply randomise between labels according to some exogenously given probability distribution $\boldsymbol{p}^0$. Each level 1 reasoner believes that his opponent reasons at level 0, and hence (if that distribution is assumed to be common knowledge – an assumption that we will reconsider later) that the opponent acts according to $\boldsymbol{p}^0$. The level 1 reasoner chooses whichever label l^* maximises his expected utility, given this belief. Each level 2 reasoner has the following beliefs about her opponent: With probability $q_0/(q_0 + q_1)$, the opponent reasons at level 0 and hence chooses according to $\boldsymbol{p}^0$. With probability $q_1/(q_0 + q_1)$, the opponent reasons at level 1, and hence chooses l^* with probability 1. The level 2 reasoner chooses whichever label l^{**} maximises her expected utility relative to these beliefs.[5] And similarly for the higher cognitive levels. It is a fundamental feature of cognitive hierarchy theory that, at each level, a player believes that his opponent's level is lower than his own; by means of this assumption, the theory generates determinate solutions rather than equilibrium conditions.

In the versions of cognitive hierarchy theory proposed by Stahl and Wilson (1995) and Camerer et al (2004), $\boldsymbol{p}^0$ is assumed to be a uniform distribution. With this assumption, the theory predicts that all strategies in a pure coordination game are chosen with equal probability. However, a theory of focal points can be generated if $\boldsymbol{p}^0$ is allowed to be non-uniform and is interpreted as describing the tendency of players to opt for the various labels when responding to them in some non-rational or non-strategic way. A very simple (but, we shall argue later, empirically inadequate) theory can be generated by assuming $\boldsymbol{p}^0$ to be common knowledge among players of level 1 or above. Given this assumption, level 0 reasoners choose according to $\boldsymbol{p}^0$ and all higher-level reasoners choose the label *l** with the highest value of p_j^0 (assuming that to be uniquely defined). If, as in Crawford and Iriberri's (2007) analysis of hide-and-seek games, the assumed properties of $\boldsymbol{p}^0$ are justified only by appeals to intuitions about salience, a theory of this kind provides a framework for organising data but has little substantive content: it predicts that players tend to choose 'salient' labels, but does not explain what 'salience' is.

One way of going further is to develop a theory of the behaviour of level 0 reasoners. Bacharach and Stahl (2000) propose a theory of this kind, which rests on strong assumptions about the formal structure of labels and about how these are perceived by players. These assumptions allow a game to be re-described in terms of the 'options' that players perceive; level 0 reasoners are assumed to choose each *option* with equal probability, but this can induce non-uniform probabilities for *strategies*.

For our purposes, however, it is not necessary to make any particular assumptions about $\boldsymbol{p}^0$. The core idea of the cognitive hierarchy approach, that focal points are induced by non-uniformities in $\boldsymbol{p}^0$, can be tested without making any prior assumptions about what those non-uniformities might be. To do this, we follow Mehta et al (1994) in defining $\boldsymbol{p}^0$ empirically, as measuring the actual frequencies with which the different labels are chosen in a *picking* task – that is, an experimental task in which players are required to 'just pick' one label from *L* in the absence of any strategic or payoff-related reasons to choose one rather than another. The tendency for a given label to be picked in such a task is its degree of *primary salience*.

This empirical definition of $\boldsymbol{p}^0$ is consistent with the logic of cognitive hierarchy theory. In those versions of the theory in which $\boldsymbol{p}^0$ is uniform, the underlying idea is that level 0 'reasoners' have no perception of *reasons for* choosing one strategy rather than another: they just pick. Because (by assumption) strategies are perceived as symmetrical with one another, each strategy is picked with equal probability. If the theory is to be generalised to allow level 0 reasoners to take account of labels in pure coordination games, it is natural to assume that such players behave as if they were facing a picking task.

It would be possible to stop at this point and specify a cognitive hierarchy model by assuming that $\boldsymbol{p}^0$, defined empirically by behaviour in a picking task, is common knowledge among reasoners of level 1 and above. But, in the context of one-shot coordination games of the kind studied by Schelling, that assumption seems implausible. For example, consider a coordination game played among university students in 2008, in which L = {«1950», «1951», ..., «2000»}. On the basis of previous research, *a theorist of focal points* might predict that, in a picking task, most respondents would pick either their own birth years or «2000» (Mehta et al, 1994). But would a typical respondent know that? And even if she did, would she be able to predict the relative frequencies of the two types of answer? And would she know the distribution of birth years among her co-players? It seems more realistic to allow for the possibility that different individuals have different beliefs about $\boldsymbol{p}^0$, and hence about which label has the greatest primary salience (that is, which label is chosen with the highest probability at level 0).

We can define a probability distribution $\boldsymbol{p}^1$ over labels such that each $p_j^{\ 1}$ is the probability that a randomly-selected player of any of the levels 1, 2, ... believes that l_j is the label with the greatest primary salience. We will say that each $p_j^{\ 1}$ is a measure of the *secondary salience* of the corresponding label l_j. Notice that $\boldsymbol{p}^1$ is not a *belief* that can be attributed to any player, or to players in general. It is a probability distribution over players' (possibly different) beliefs about primary salience.

What relationship should we expect to find between $\boldsymbol{p}^1$ and $\boldsymbol{p}^0$? Consider any player i of level 1 or above. By imagining herself 'just picking', she can simulate the behaviour of a level 0 reasoner. If this simulation of picking is not simply the application of a random device, and if it is governed by the same mental process as governs actual picking, the result of the simulation – the simulated pick of a particular strategy – provides i with some information about the behaviour of level 0 reasoners. If this were the *only* relevant information available to i, the label that she picked would also be the label that she believed to have the greatest primary salience (i.e. the label that she believed to be modal in $\boldsymbol{p}^0$). If this were true for all players, we would have $\boldsymbol{p}^1 = \boldsymbol{p}^0$. But in coordination games of the kind described by Schelling, in which labels have distinct and meaningful descriptions in terms of the players' own language, culture or experience, it is reasonable to suppose that some players *do* have additional information. Games in which labels are meaningful in this sense will be called *describable*.

To understand the nature of this information, take the case of the game in which L = {«1950», «1951», ..., «2000»}. Consider a player who happens to be considerably older than most of her fellow-students. She imagines herself picking, and picks «1973». On reflection, she realises that the special feature of this label is that she was born in 1973.

Combining the information that she has picked her birth year with her imperfect background knowledge of the age distribution of university students, she might form the belief that the mode of $\boldsymbol{p}^0$ is the modal birth year of current students, and that this is 1988. If everyone behaves in this way (that is, picking her own birth year but attributing greatest primary salience to the year she believes to be the most common birth year) and if errors are random, $\boldsymbol{p}^0$ will have same distribution as actual birth years; the distribution of $\boldsymbol{p}^1$ will have the same mode, but will be less dispersed. Generalising from this example, whenever players have some understanding of their own propensities to pick some labels rather than others, we should expect that $\boldsymbol{p}^1$ and $\boldsymbol{p}^0$ have the same mode and that $\boldsymbol{p}^1$ is less dispersed.

In principle, this line of analysis could be extended indefinitely. The next step would be to ask what players of levels 2 and above believe about $\boldsymbol{p}^1$. Just as in the case of $\boldsymbol{p}^0$, it is implausible to assume that $\boldsymbol{p}^1$ is common knowledge among these higher-level reasoners. We might define a probability distribution $\boldsymbol{p}^2$ over labels such that each p_j^2 is the probability that a randomly-selected player of any of the levels 2, 3, ... believes that l_j is the label with the greatest secondary salience; each p_j^2 measures the *third-order salience* of the corresponding label l_j. And so on.

For typical experimental applications, however, it seems unlikely that third- and higher-order salience differ from secondary salience. Consider again the mature student in the coordination game with $L = \{$«1950», «1951», ..., «2000»$\}$. Her own simulated pick is «1973»; she believes that other subjects pick their birth years and that the modal birth year is 1988; so she predicts (or 'guesses') that the modal pick is «1988». We are now asking what prediction she would make about the modal *guess* of other subjects. If she attributes to them the same mode of reasoning that she used in her own guessing, she will predict that (except for random error) they all guess «1988». *In principle*, it is conceivable that she has background knowledge about the reasoning process by which subjects make such guesses, and that she knows that her own reasoning about this matter is atypical; but such knowledge is far more esoteric than the analogous knowledge about picking. It seems reasonable to assume that subjects do *not* have such information, and hence that each subject's prediction of the modal guess of other subjects is the same as her own guess. This implies $\boldsymbol{p}^1 = \boldsymbol{p}^2$. Repeating this argument for successively higher levels of reasoning, we have $\boldsymbol{p}^1 = \boldsymbol{p}^2 = \boldsymbol{p}^3 = \dots$. In other words, the behaviour of all players of levels 1, 2, … is predicted by $\boldsymbol{p}^1$.

If this result holds, behaviour in a coordination game can be predicted using the information generated by a *guessing* task with the following structure: individuals are asked to guess which label from L was chosen by an unknown other subject in a picking task, and they are rewarded for guessing correctly. The inclusion of a guessing treatment is one of the key innovations of the experiments reported in this paper.

For players of level 1 or above, the guessing task can be interpreted as asking for a judgement about which label is primarily salient (that is, modal in $\boldsymbol{p}^0$); thus, we should expect the responses of such players to be predicted by $\boldsymbol{p}^1$. Given the logic of the cognitive hierarchy hypothesis, with its implicit assumption that level 0 reasoners do not consider how their opponents might behave, it is natural to assume that, when guessing, such individuals merely report what they themselves would have chosen in the picking treatment, and hence that their responses are predicted by $\boldsymbol{p}^0$. But these assumptions imply a distribution of responses to the guessing task that is exactly the same as the distribution predicted for the coordination game. (Consider a randomly-selected player in the coordination game. With probability $1 - q_0$ she reasons at level 1 or above, and so her behaviour is predicted by $\boldsymbol{p}^1$; with probability q_0 she is a level 0 reasoner, and so her behaviour is predicted by $\boldsymbol{p}^0$.)

Summing up, the preceding analysis has generated two hypotheses about pure coordination ('PC') games which can be tested in an experiment with counterbalanced picking, guessing and coordination treatments:

Hypothesis PC1: In any pure coordination game, the distribution of responses is at least as concentrated for guessers as it is for pickers. If the game is describable, the distribution is more concentrated for guessers.

Hypothesis PC2: In any pure coordination game, the distribution of responses is the same for coordinators as for guessers.

These hypotheses decompose Mehta et al's finding that coordinators' responses are more concentrated than pickers'. Hypothesis PC1 is implied by plausible assumptions about guessers' information on the factors that influence pickers; it is not specific to cognitive hierarchy theory. In contrast, hypothesis PC2 is a distinctive implication of cognitive hierarchy theory, and provides a test of that theory's explanation of focal points. Notice that these hypotheses do not depend on any prior assumptions about the relative salience of different labels. Thus, tests of these hypotheses can use coordination games in which salience is subjective and culturally dependent.

1.3 *Schelling salience in pure coordination games: the theory of team reasoning*

We use the term *Schelling salience* in the same sense as Mehta et al, who quote Schelling's explanation of why, in a pure coordination game with the instruction 'Name a positive number', the number 1 is the modal choice, even though it is not the most common response when people are asked just to pick a number. Schelling says: 'If one ... asks what number, among all positive numbers, is most clearly unique, or *what rule of selection would lead to unambiguous results*, one may be struck by the fact that the universe of all positive numbers has a "first" or "smallest" number' (1960, p. 94, italics in original). The implication is that a 'rule of selection' is a criterion that a player can use to choose a label from the relevant set *L*; in the case of choosing from a set of integers, examples might include 'Choose the smallest number', 'Choose your favourite number', 'Choose the number with the largest number of prime factors' and so on. As these examples suggest, rules differ both in their probability of being recognised and (given that they are recognised and followed by both players) in their probability of leading to coordination. Schelling's idea seems to be that the players look for a rule which clearly outperforms its rivals on these criteria; such a rule has 'Schelling salience'.

This idea has been developed using the concept of *team reasoning* by Sugden (1993, 1995) and Michael Bacharach (1999, 2006). An individual *i* team-reasons with respect to a group *G* if she works out which profile of options for members of *G* would give the best results for *G*, and then chooses her component of that *team-optimal* profile. Roughly, the individual asks 'What should we do?', and acts upon the answer in the expectation that other members of the group think and behave analogously. Whether the players of a particular game actually use team reasoning may depend on the nature of the game. Bacharach (2006) proposes that coordination games are particularly likely to prompt team reasoning, because the players' interests are aligned and there are opportunities for mutual gain.

If the profiles of 'options' over which players optimise are interpreted as *strategy* profiles, the team-reasoning hypothesis has the same implications for behaviour in coordination games as the hypothesis that players use payoff dominance as an equilibrium selection device, as proposed by Harsanyi and Selten (1988).[6] One of those implications is that, in pure coordination games, strategies are chosen at random. If the theory of team reasoning is to explain focal points in pure coordination games, an 'option' must be interpreted similarly to a ‘rule of selection’ in Schelling’s analysis, with no requirement of a one-to-one correspondence between options and strategies.

Some theorists have followed the approach that Bacharach and Stahl (2000) use in conjunction with cognitive hierarchy theory – that is, to use assumptions about the formal structure of labels to re-describe games in terms of the options that the players themselves perceive, and then to assume that players optimise over profiles of such options (Bacharach, 1999, 2006; Casajus, 2001; Janssen, 2001).[7] Another approach is to assume that players observe independent realisations of some payoff-irrelevant process which, with non-uniform probabilities, picks out (or ‘mentions’) labels from the set *L*; this allows players to use rules of selection such as ‘Choose the most-frequently mentioned label’ (Sugden, 1995).

For our purposes, however, there is no need to presuppose a particular formal model of rule selection because, as we now explain, our design will allow us to test more general implications of the team reasoning approach by comparing responses to guessing and coordination treatments.

We start from the observation that primary and secondary salience can themselves be used as rules of selection. ‘Choose a label as if you were just picking’ (or ‘Choose the label with the greatest immediate appeal to you’) seems a credible rule of selection, and corresponds with primary salience. ‘Choose the label most likely to be picked by someone who is just picking’ (or ‘Choose the label most likely to have immediate appeal to an average person’), is equally credible, and corresponds with secondary salience. For experimental subjects confronting pure coordination games for the first time, the sheer oddness of having to choose from a set of apparently arbitrary labels seems likely to cue thoughts about just picking. Thus, one might expect subjects who are capable of team reasoning to be aware of these two rules. For the reasons explained in section 1.2, two co-players will generally have a greater probability of coordinating if they both follow the secondary salience rule than if they both follow the primary salience rule. Hence, if a team-reasoning player cannot find a rule of selection which gives a higher probability of coordination than the secondary salience rule, she will follow the latter rule. In this case, secondary salience and Schelling salience coincide.

Thus, the theory of team reasoning is not disconfirmed if, as predicted by cognitive hierarchy theory, guessing and coordination treatments generate the same distribution of responses. But if the two distributions are different, we can ask whether the differences have the characteristics that would be expected, were the theory of team reasoning correct. Since team reasoners look for a team-optimal rule of selection, they should reject the secondary salience rule only in favour of rules which give at least as great a probability of coordination. If guessing and coordination treatments generate different distributions of responses, the distribution from the latter treatment should be at least as concentrated as that from the former. Thus, the theory of team reasoning implies the following hypothesis:

> *Hypothesis PC3: In any pure coordination game, if the guessing and coordination treatments generate different distributions of responses, the distribution from the coordination treatment is at least as concentrated as that from the guessing treatment.*

1.4 *Nondescript Hi-Lo games*

The principles underlying the two rival hypotheses can be tested in another way, by adapting an example discussed by Schelling (1960, pp. 295-296). Schelling considers a Hi-Lo game with $n = 4$, $U_1 = 10$, $U_2 = 10$, $U_3 = 10$, and $U_4 = 9$. If we consider only the normal form of the game, there are three completely symmetrical pure-strategy Nash equilibria and one further such equilibrium, distinguished from the others by giving a *lower* payoff to both players. Schelling asks us to assume that 'the strategies occur in a way that makes ordering them intellectually impossible for rational players'. In our framework, in which it is a matter of definition that every strategy has a unique label, the closest approximation to Schelling's assumption is to make the differences between the labels *nondescript* – that is, such that, although normal players are aware that the labels are not the same, they do not have any readily-available way of describing those differences, even to themselves.[8] If all the labels in a coordination game are nondescript, we will say that the game itself is 'nondescript'. Schelling claims of his game: '[I]f no better means of coordination can be discerned, the "solution" may be the strategy pair ... with payoffs of 9 apiece'.

This conclusion follows from a straightforward extension of the team-reasoning analysis in section 1.3. Because the labels are nondescript, there is no obvious rule which unambiguously picks out one of the *labels* by virtue of its standing out. However, there is an apparently obvious rule which, if followed by both players, would lead them both to choose l_4 by virtue of the corresponding *payoff* . This is the rule 'Choose the label attached to the payoff that is the odd one out'. The opposite rule, 'Pick one of the labels attached to the highest payoff', is sub-optimal in the team-reasoning sense (on the assumption that players seek jointly to maximise expected utility). Thus, the hypothesis of team reasoning implies that l_4 is chosen.

In contrast, consider the implications of cognitive hierarchy theory. The first step in applying this theory is to specify $\boldsymbol{p}^0$, the distribution of the responses of level 0 reasoners. Recall that level 0 reasoners are people who do not engage in any kind of strategic reasoning; they act as if unaware that they are interacting with anyone. One possible assumption is that these individuals are completely unaware of the significance of payoffs, and so just pick among labels (as, in our analysis, they do in pure coordination games). If the labels are nondescript, this is equivalent to picking at random. Given that level 0 reasoners can be expected to behave in this way, level 1 reasoners are indifferent between l_1, l_2 and l_3 (each of which they believe will give an expected utility of 10/4) but strictly prefer each of these to l_4 (which they believe will give 9/4). If level 1 reasoners randomise between l_1, l_2 and l_3, level 2 reasoners are also indifferent between these three labels and prefer each of them to l_4, and so on. The overall implication is that l_4 is chosen with probability $q_0/4$, while each other label is chosen with probability $q_0/4 + (1 - q_0)/3$.

Alternatively, in specifying $\boldsymbol{p}^0$, we might assume that level 0 reasoners take some account of payoffs, but in a non-strategic way. It is natural to assume that, for a player who is not thinking strategically, higher payoffs have a stronger tendency to prompt positive affective responses than lower payoffs do – in the same sense that, in a picking task, «Porsche» is a more attractive label than «Volkswagen». If, as we conjecture, primary salience is associated with pre-reflective attractiveness, level 0 reasoners will choose l_4 with probability less than 1/4 and higher-level reasoners will not choose it at all, with the result that its overall probability of being chosen is less than $q_0/4$.[9]

This analysis can be extended to the general class of nondescript Hi-Lo games. Consider any such game. Suppose that there are n_1 labels for which the payoff is x_1, n_2 labels for which the payoff is x_2, ... , and n_m labels for which the payoff is x_m, where $x_1 > x_2 > ... > x_m$. Then the rule 'Pick one of the labels associated with a payoff of x_k', if followed by both players, would give each an expected payoff of x_k/n_k. Each of the n_k labels associated with the k that maximises the value of x_k/n_k is *team-optimal*. Team reasoning requires each player to pick from the set of team-optimal labels. For example, in a game in which there are six labels associated with payoffs 10, 10, 10, 9, 8, 7, the optimal rule is to choose the label with the payoff 9; in a game in which there are five labels and payoffs 10, 10, 10, 10, 1, the optimal rule is to pick from the set of labels with payoff 10 (giving an expected payoff of 2.5). In contrast, cognitive hierarchy theory implies that every player of level 1 or above randomises among the labels associated with the highest payoff, while level 0 players randomise among all labels (possibly giving greater weight to labels associated with higher payoffs). Thus, averaging across players of all levels, the choice probability for each of the labels associated with the highest payoff x_1 is at least $q_0/n + (1 - q_0)/n_1$.

The analysis in the preceding paragraph assumes that players of level 1 or above maximise expected utility (and that this is common knowledge) and that *utility* payoffs are common knowledge. In applying this analysis to games in which payoffs are described in material units such as money, some allowance must be made for players' attitudes to risk, and for these attitudes not being common knowledge. However, it seems reasonable to assume it to be common knowledge that players' attitudes to risk are not pathologically distant from risk neutrality. Thus, in the first example discussed in the preceding paragraph, if payoffs are in British pounds, it seems uncontroversial to assume that the certainty of £9 is preferred to a 0.33 chance of £10. In the second example, it is probably safe to assume that a 0.25 chance of £10 is preferred to the certainty of £1.

Summing up, we have generated two rival hypotheses about behaviour in the coordination treatment of Hi-Lo ('HL') games. (The formulations we use below allow for random error in players' choices.) Hypothesis HL1 is implied by cognitive hierarchy theory, while HL2 is implied by team reasoning:

> *Hypothesis HL1: In any nondescript Hi-Lo game, the choice probability for each of the labels associated with the highest payoff is greater than that for every label associated with a lower payoff.*
>
> *Hypothesis HL2: In any nondescript Hi-Lo game, the choice probability for each team-optimal label is greater than that for every other label.*

2 Experimental Design

2.1 *Features common to both experiments*

We implemented two versions of the same design, conducted in March 2001 using subjects recruited from the general student populations of the University of Amsterdam in the Netherlands (for one experiment) and the University of Nottingham in the UK (for the other).[10] In each case, subjects faced a series of tasks. In each task, the subject was presented with a set of objects and was required to choose one. Each object was associated with a specified number of points. There were three treatments. In the *picking* treatment, the subject was simply asked to choose one object, and scored the number of points specified for that object. In the *guessing* treatment, the subject was paired with a randomly-selected anonymous partner in the picking treatment, and was asked to guess which object her partner had chosen; this pairing was the same for all tasks. If this guess was correct, the guesser scored the number of points associated with the relevant object; otherwise, she scored nothing. In the *coordination* treatment, the subject was paired with a randomly-selected anonymous partner facing the same task in the same treatment; again, the pairing was the same for all tasks. If the two partners chose the same object, both scored the number of points associated with it; otherwise, both scored nothing. In all treatments, subjects were unable to communicate with one another. No feedback was given until the end of the experiment, when subjects were paid in proportion to the total number of points scored.

In implementing these three treatments, we used tasks of two types. In a *text task*, all objects carry the same number of points (10 in all such tasks in both experiments), but each has a distinct label in the form of a string of text. For example, one text task in the Amsterdam experiment contains four objects with the labels «Jaguar», «Ford», «Porsche», «Ferrari». In a *number task*, the objects may carry different numbers of points, but in other respects they are (as far as possible) nondescript. Notice that, when presented in the coordination treatment, text tasks are describable pure coordination games. Number tasks in which all objects carry the same number of points are nondescript pure coordination games. Other number tasks are nondescript Hi-Lo games.[11]

Subjects were allocated at random between coordination and picking/guessing sessions. In the coordination sessions, all subjects faced the whole set of text and number tasks in the coordination treatment. In picking/guessing sessions, subjects first faced half of the set of text tasks and half of the set of number tasks in the picking treatment. They then faced the remaining tasks in the guessing treatment. Within each treatment, the order in which tasks were presented to subjects was randomised. The design was counterbalanced so that each task was faced in each of the three treatments, in each case by a different set of subjects. Subjects were allocated to sessions so as to generate approximately equal numbers of responses for the three treatments.

Because picking was always done before guessing, and because the instructions for the guessing tasks were not given until the picking tasks had been completed, pickers had no reason to think of their responses as having any effect on other subjects. However, we hoped that guessers' prior experience of picking would help them to understand the picking task that their partners had faced. Because coordinators were not aware of the picking and guessing tasks, reasoning in the coordination treatment could not be cued by ideas suggested by the other two treatments.

In the Amsterdam experiment, 164 subjects were randomly allocated to 15 sequential sessions. Three observations were lost through computer crashes, resulting in sample sizes of 53, 52 and 56 subjects for the picking, guessing and coordination treatments respectively. Subjects were paid at a pre-announced rate of 15 Dutch cents per point ($0.06 at the exchange rate of the time), in addition to a fixed show-up fee of 5 guilders ($2.11); average earnings were 30.12 guilders ($12.05) per subject. In the Nottingham experiment, 134 subjects took part in three simultaneous sessions, resulting in sample sizes of 45, 45 and 44 for the three treatments. Subjects were told at the start of the experiment that payment would be at a constant rate per point, to be calculated ex post to ensure an average payment of £7 ($10.43) per subject for the experiment as a whole.

Although the two experiments shared a common basic design, there were some differences in the presentation of tasks to subjects, and different sets of tasks were used. These features of the experiments are described in the following two sections.

2.2 *Presentation of tasks*

The Amsterdam experiment was computerised. The objects from which a choice had to be made were presented as discs moving within a rectangular field on the computer screen. Each disc moved in a straight line until it collided with a border or another disc, in which case it rebounded in a randomly perturbed direction. The subject selected a disc by clicking on it with the mouse. In a text task, both the relevant piece of text and the number of points was written on each disc. In a number task, only the number of points was shown. Figures 1 and 2 show examples of the two types of task, as represented in this display. (Lines have been added to indicate movement; these did not appear in the experiment.) Since the pattern of movement of the discs in any given task was the same for both members of any given pair of co-players, the movement of each disc gave it a distinct label, even in a number task in which two or more discs carried the same number of points. (To avoid confounds, patterns of movement were varied across sessions.) We expected that this kind of labelling would be perceived as nondescript by most subjects, while text differences would be immediately obvious.

Instructions were given both orally (to all subjects in the session together, to ensure common knowledge) and on subjects' computer screens. The relevant instructions in the picking treatment (described to subjects as 'part 1' of the experiment) were:

> In this part of the experiment, your earnings are determined by your decisions alone. There are fourteen tasks in part 1. Each task shows a set of moving objects, with a number on each one. The display in each task can be thought of as a short 'film'. For each task, you have to click on one object in each film, using your mouse. … For that task, you will earn the number of points shown on the object.

The corresponding instructions in the guessing treatment were:

> Each 'film' in part 2 is one that your partner had during part 1, in which he or she just clicked on an object and received the number of points written on it. Again, you have to click on one object for each task, and confirm your decision. This time, though, you have to guess what your partner did during part 1. If you click on the same object as your partner, you will receive the number of points indicated on that object. If not, you will receive nothing for that task.

In the coordination treatment, the instructions were:

> Each task shows a set of moving objects, with a number on each one. The display in each task can be thought of as a short 'film'. Your partner has the same set of films. For each task, you have to click on one object in each film, using your mouse. … If you click on the same object as your partner, you will both receive the number of points indicated on that object. If not, neither of you will receive anything for that task.

In the Nottingham experiment, tasks were presented in booklets. Each task appeared as a row of five objects, and the subject selected one by marking a tick below it. Subjects who had been paired with one another saw the same five objects, but the order in which these were displayed from left to right was randomised across subjects (and subjects were told this). Each object was represented as a box, subdivided into two parts. The lower part stated the number of points associated with the object. In text tasks, the upper part of each box contained a distinct string of text; figure 3 shows a typical example. In number tasks, the upper part of each box contained a distinct pattern of symbols; figure 4 shows an example. These five patterns were generated by separate runs of a common computer program which included a random component. Patterns were generated independently for each pair of subjects. Our intention was that these patterns, although clearly constituting distinct labels, would be perceived as nondescript.

The relevant instructions (given both orally, to all participants together, and in print) were:

> [Picking treatment] Your objective is the same for each task: *to pick one of the boxes*. You are required to indicate which box you have chosen by putting a tick just below the box. ... For each of the sixteen tasks, you will be awarded the number of points specified in the box you have picked. The total number of points awarded to you for all the tasks determines how much money you win in this part of the experiment.
>
> [Guessing treatment] There is an even number of people taking part in this room, and we have randomly divided you into pairs for the duration of this part of the experiment. .... What you see in your second booklet is the same as your partner saw in their first booklet when you were all asked to pick one of the five boxes for each task. So for each task in your second booklet, your partner has already chosen one of the five boxes and scored the corresponding number of points, which they keep regardless of what you do next. Your objective for each task now is: *to guess which of the boxes your unknown partner picked*. You are required to indicate which box you think this is by putting a tick just below the box. ... If you correctly guess which box your partner picked, then you will be awarded the number of points specified in the box. If you fail to guess which box your partner picked, you will not receive any points for that task. The total number of points awarded to you for these sixteen tasks determines how much money you win in this part of the experiment.
>
> [Coordination treatment] There is an even number of people taking part in this room, and we have randomly divided you into pairs for the duration of the experiment. ....Your objective is the same for each task: *to choose the same box as that of your unknown partner*. You are required to indicate which box you have chosen by putting a tick just below the box. ... If the pair of you choose the same box, then you as an individual will be awarded the number of points specified in the box. If the pair of you fail to choose the same box, you will not receive any points for that task. The total number of points awarded to you for all the tasks determines how much money you win.

2.3 *Text tasks*

Each experiment used fourteen text tasks, denoted TA1–TA14 (for Amsterdam) and TN1–TN14 (for Nottingham). These tasks used the following sets of labels (the string symbols « and » are omitted to reduce clutter):

Amsterdam text tasks[12]

TA1: {grijs, indigo, karmozijn, magenta, turkoois}
TA2: {Ferrari, Ford, Jaguar, Porsche}
TA3: {Berlin, Brussel, Lissabon, Madrid, Mannheim}
TA4: {almond, cashew, peanut, walnut}
TA5: {diamond, emerald, glass, sapphire}
TA6: {chrome, copper, iron, plastic, steel}
TA7: {bread, curry, pizza, steak}
TA8: {beer, sherry, water, whisky, wine}
TA9: {Carlsberg, Corsendonk, Grimbergen, Rochefort, Westmalle}[13]
TA10: {frog, leopard, panther, tiger}
TA11: {aeroplane, bicycle, helicopter, hovercraft}
TA12: {chess, football, squash, tennis, volleyball}
TA13: {Barbados, Bern, Florida, Honolulu}
TA14: {jogging, running, sitting, walking}

Nottingham text tasks

TN1: {Friday lunchtime, Monday morning, Saturday night, Sunday night, Wednesday evening}
TN2: {Earth, Mars, Mercury, Saturn, Venus}
TN3: {Ford, Mercedes, Pontiac, Porsche, Volkswagen}
TN4: {cheese omelette, ham omelette, mushroom omelette, plain omelette, prawn omelette}
TN5: {1, 2, 7, 10, 15}
TN6: {deck chair, dining chair, easy chair, rocking chair, stool}
TN7: {Colorado, Florida, Louisiana, Nevada, Ontario}
TN8: {jogging, sitting, sunbathing, swimming, walking}
TN9: {1978, 1979, 1980, 1981, 2000}
TN10: {David, John, Michael, Robert, Steven}
TN11: {win champagne, win chocolate, win money, win nothing, win trophy}
TN12: {blue, green, orange, purple, red}
TN13: {apple juice, carrot juice, grapefruit juice, mango juice, pineapple juice}
TN14: {Berlin, Calais, Paris, Prague, Rome}

In composing these sets of labels, we tried to ensure that Schelling salience and secondary salience would diverge, so as to increase the potential for team reasoning, if operative, to generate differences between the responses of guessers and coordinators. However, we emphasise again that our formal hypothesis tests apply to *any* set of labels; they are not conditional on any particular characteristics of our tasks.

Our aim (which, as will emerge later, we achieved with varying degrees of success) was that in each task, one of the labels, say l_1, would be unambiguously picked out by some obvious rule of selection other than primary or secondary salience; we will call this label the *intended salient*. In the Amsterdam tasks, the relevant rule was always ‘Choose the odd one out’. The Nottingham tasks were composed with the intention that each of them would evoke one or other of the rules that seemed to have been used by subjects in Mehta et al's coordination treatment; in addition to 'Choose the odd one out', these included 'Choose the archetype' and 'Choose the status quo'.[14] We intended that each of the other labels l_2, .., l_n should be roughly equal in the kind of immediate appeal which is likely to induce primary salience, while l_1 should have either the same or less appeal.

For example, consider the set of labels used in TA3: {«Berlin», «Brussel», «Lisbon», «Madrid», «Mannheim»}. Here, «Mannheim» is the odd one out: all the other cities are national capitals. We conjecture that immediate appeal is determined by a person's affective response to whatever ideas are suggested by the labels. In this case, one might expect each of the four capital cities to evoke ideas of national or cultural significance, or of attractiveness as a tourist destination; for any individual, the relative force of these ideas would depend on matters of taste, culture, nationality and personal association. By comparison, Mannheim is not generally credited with comparable positive qualities. Thus, one might expect $\boldsymbol{p}^0$ to have a dispersed distribution. Since people will find it difficult to judge which label is modal in $\boldsymbol{p}^0$, the distribution of $\boldsymbol{p}^1$ will be dispersed too, and so co-players who follow the rule of secondary salience will be relatively unsuccessful. In particular, because «Mannheim» is so obviously the odd one out in L, they will be less successful than they would be by following the rule 'Choose the odd one out'.

To allow readers to test their own intuitions, the intended salient for each task is identified only in a footnote.[15] If the reader's intuitions sometimes differ from ours, he or she should remember that the intended salient plays no role in our hypothesis tests.

2.4 *Number tasks*

The Amsterdam experiment included fourteen number tasks, NA1–NA14. The Nottingham experiment included eighteen such tasks, NN1–NN18. For our purposes, the main characteristic of a number task is the array of points carried by the set of objects from which the subject must choose. The following arrays were used:

Amsterdam number tasks

Type 1
NA1: (10, 10, 10, **9**)
NA2: (10, 10, 10, 10, 10, **9**)
NA3: (10, 10, 10, **9**, 8, 7)

NA4: (10, 10, 10, 9, 9, **8**)
NA5: (10, 10, 10, 10, **9**, **9**)

Type 2
NA6: (**10**, 9)
NA7: (**10**, **10**, **10**, 9, 9, 9)
NA8: (**10**, 1)
NA9: (**10**, **10**, **10**, 1)
NA10: (**10**, **10**, **10**, **10**, **10**, 1)

Type 3
NA11: (**10**, **10**)
NA12: (**10**, **10**, **10**, **10**)
NA13: (**10**, **10**, **10**, **10**, **10**)
NA14: (**10**, **10**, **10**, **10**, **10**, **10**)

Nottingham number tasks

Type 1
NN1, NN2, NN3, NN4, NN5, NN6: (10, 10, 10, 10, **9**)

Type 2
NN7, NN8, NN9, NN10, NN11, NN12: (**10**, **10**, **10**, **10**, 1)

Type 3
NN13, NN14, NN15, NN16, NN17, NN18: (**10**, **10**, **10**, **10**, **10**)

Coordination-treatment responses that are consistent with the theory of team reasoning (under credible assumptions about risk attitudes)[16] are shown in bold. In all cases, the cognitive hierarchy hypothesis implies that subjects who reason at level 1 or above choose 10-point options.

These tasks are divided into three types. In type 1 tasks, the coordination treatment is a nondescript Hi-Lo game in which the team reasoning hypothesis implies that players will choose an option which does *not* carry 10 points. These tasks allow a direct comparison between the two hypotheses. In type 2 tasks, the coordination treatment is a nondescript Hi-Lo game in which both hypotheses have the same implications for behaviour. These tasks are significant because they subject the team reasoning hypothesis to an additional test. In almost all type 1 tasks, the team-optimal response is also the option with the *lowest* number of points, and, in terms of points, is the odd one out. In type 2 tasks, however, choosing options with these characteristics is contrary to the team reasoning hypothesis. In type 3 tasks, all options carry 10 points, and so the coordination treatment is a pure coordination game. These tasks allow us to test the background assumption that the labels associated with the options are nondescript. If that assumption were true, the NCI for the coordination treatment of a type 3 task would equal 1 (plus or minus random noise).

In the Nottingham experiment, each of three arrays of points (one for each of the task types) occurs in six different tasks. Recall that each number task involved a set of five randomly-generated patterns. We used six different pattern-generating programs; each program was paired with each array of points in a factorial design. This allowed us to investigate whether subjects' responses were influenced by the kinds of patterns they were shown. In fact, we found no pattern-specific effects.

3 Results for text tasks

3.1 *Presentation of results*

Table 1 reports the frequency distribution of responses for each of the 28 text tasks and for each of the three treatments. For each distribution, the NCI is shown at the bottom of the relevant column. For each task, table 1 reports four tests.

The first of these tests whether, as predicted by hypothesis PC1, guessers' responses are more concentrated than those of pickers. We use a bootstrap method (Efron 1979). We start with the null hypothesis that guessers' responses are drawn from a distribution with the same relative frequencies as the actual responses of pickers. We obtain critical values of the NCI that would be generated by repeated sampling, with a sample size equal to the actual number of guessers, if the null hypothesis were true. Our estimates of these critical values are constructed from 20 000 simulated samples. For example, consider task TA1. The NCI is 0.935 for pickers and 1.040 for guessers. In 5 per cent of our simulations, the NCI for guessers is greater than 1.130. Since 1.040 < 1.130, we cannot reject the null hypothesis at the 5 per cent level in a one-tail test. This finding, that the NCI is not significantly greater for guessers than for pickers, is reported by the entry 'ns' in the 'guess' column against 'significance wrt pick'. Cases in which the null hypothesis can be rejected at the 5 per cent level or 1 per cent level are reported as * or **; cases in which the observed difference is in the 'wrong' direction are recorded as #. Using the same method, we test whether, as in the data reported by Mehta et al, coordinators' responses are more concentrated than pickers'. The results of these tests are reported against 'significance wrt pick' in the 'coordinate' column. And, most importantly, we test whether coordinators' responses are more concentrated than guessers'; the results of these tests are reported against 'significance wrt guess' in the 'coordination' column. Recall that hypothesis PC2, implied by cognitive hierarchy theory, predicts that the two distributions are the same (and hence equally concentrated), while PC3, implied by team reasoning, predicts that if the two distributions are different, coordinators' responses are at least as concentrated as guessers'.

Finally, we report a chi-squared test of the null hypothesis that coordinators' and guessers' responses are drawn from the same population distribution. The result of this test is reported against 'chisq wrt guess' in the 'coordination' column; rejection of the null at the 5 per cent or 1 per cent level is denoted by * or **.[17] This is a direct test of PC2.

For some purposes, it is convenient to work with summary statistics which aggregate across tasks. For each experiment, table 2 reports five such statistics concerning pickers, guessers and coordinators. (The entries in the other rows and columns will be explained later.) The entry in the (pick, pick) cell is the NCI for pickers (denoted $\mathrm{NCI}_{\mathrm{PP}}$), averaged across the relevant experiment.[18] The entries in the (guess, guess) and (coordinate, coordinate) cells are the corresponding average NCIs for guessers and coordinators, denoted $\mathrm{NCI}_{\mathrm{GG}}$ and $\mathrm{NCI}_{\mathrm{CC}}$. The other entries are averages of *cross-group NCIs*, defined as follows. Consider a label set $L = \{l_1, ..., l_n\}$, and two disjoint groups of individuals, one with N members and one with N'members. Each individual chooses one label from L. Each label l_j is chosen by m_j individuals from the first group and by m_j' from the second. The *cross-group coordination index*, $\sum_j m_j m_j'/NN'$, measures the probability that an individual drawn at random from one group will choose the same label as an individual drawn at random from the other. Since this index takes the value $1/n$ when individuals in one or both groups choose at random, we can multiply it by n to arrive at the *cross-group NCI*. Cross-group NCIs provide information about how far the responses of subjects in different treatments are concentrated *on the same labels*.

3.2 *Amsterdam results*

We begin by looking at the aggregated data, shown in table 2. Notice that $\mathrm{NCI}_{\mathrm{PP}} = 1.116$, $\mathrm{NCI}_{\mathrm{GG}} = 1.259$, and $\mathrm{NCI}_{\mathrm{CC}} = 1.819$. That is, relative to the benchmark case of subjects who choose at random, a pair of pickers is only 12 per cent more likely to give matched responses, while the corresponding figures for guessers and coordinators are 26 per cent and

82 per cent respectively. The marked difference between the concentration of pickers' and coordinators' responses replicates the main finding of Mehta et al. That guessers' responses are more concentrated than pickers' is consistent with PC1. That coordinators' responses are much more concentrated than guessers' is contrary to PC2 (and cognitive hierarchy theory) but consistent with PC3 (and team reasoning). Notice also that $NCI_{PC} \approx NCI_{PG}$ and $NCI_{GC} \approx NCI_{PG}$: guessers are almost as successful as coordinators in matching the responses of both pickers and (other) guessers. The implication is that coordinators' differential success in matching one another is not the result of their responses being highly concentrated on the modal responses of pickers and/or guessers, which again is suggestive of team reasoning.

We now consider the data for individual tasks, shown in table 1. Guessers' responses are more concentrated than pickers' in twelve of the fourteen tasks, as predicted by PC1; in eleven of these tasks, the difference between the two NCIs is statistically significant. The distributions of responses of coordinators and guessers are significantly different from one another, contrary to PC2 and cognitive hierarchy theory, in thirteen out of fourteen tasks. These thirteen cases can be used for tests of PC3. In all thirteen cases, consistently with PC3 and team reasoning, $NCI_{CC} > NCI_{GG}$; the difference is statistically significant in nine cases. Modal choices are different in seven cases, which again is suggestive of team reasoning. In twelve of the fourteen tasks, the modal choice of coordinators is the intended salient. (The exceptions are «almond» in TA4 and «curry» in TA7.)

Taken together, these findings suggest that coordinators are using a mode of reasoning which is different from that used by guessers, and which generates responses which are both more concentrated and more skewed in favour of the 'odd one out'. We interpret all this as strong evidence against cognitive hierarchy theory, and as supportive of the theory of team reasoning.

3.3 *Nottingham results*

Aggregating across tasks, $NCI_{PP} = 1.204$, $NCI_{GG} = 1.982$, and $NCI_{CC} = 2.197$ (see table 2). Again, Mehta et al's main result is replicated. Guessers' responses are *much* more concentrated than pickers', strongly supporting PC1. In contrast to the Amsterdam results, however, coordinators' responses are only slightly more concentrated than guessers. In none of the fourteen tasks is there a significant difference between the distributions of responses for coordinators and guessers. These findings are entirely consistent with PC2 and cognitive hierarchy theory. Since there is no evidence of systematic differences between the responses of coordinators and guessers, PC3 does not apply. It seems that the Nottingham coordinators are responding as if they were guessers.

4 Results for number tasks

4.1 *Presentation of results*

Table 3 reports the frequency distribution of responses for each of the type 1 and type 2 number tasks, for each of the three treatments. Responses are classified only by the number of points carried by the options chosen. For example, task NA1 has four options with the array of points (10, 10, 10, 9), but we disaggregate responses into '10-point options' (denoted '10x3' to signify that there are three options, each carrying 10 points) and '9-point options' (denoted '9x1'). Our designs preclude further disaggregation. As explained in section 2.2, the procedures by which individual options were labelled (the 'films' in the Amsterdam experiment and the 'patterns' in the Nottingham experiment) were randomised between sessions (in Amsterdam) or between subject pairs (in Nottingham).

For each task and each treatment, we report a one-tail binomial test of the hypothesis that 10-point options are chosen with greater probability than if choice were random (that is, that this probability is greater than the proportion of options which carry 10 points). For the coordination treatment of type 1 tasks, this is a test of HL1, which is implied by cognitive hierarchy theory. For the coordination treatment of type 2 tasks, it is a test of both HL1 and HL2, and so does not discriminate between the two theories. For the coordination treatment of type 1 tasks, we report a corresponding test of whether team-optimal options are chosen with greater probability than if choice is random. This is a test of HL2, which is implied by team reasoning. Cases in which the null hypothesis is rejected at the 5 per cent (1 per cent) level are denoted by * (**) in the final column of table 3. Cases in which the options relevant for the test are chosen with *lower* frequency than would be implied by random choice are denoted by #.

Our tests of team reasoning are premised on the assumption that, in number tasks, subjects perceive labels as nondescript. One check on the validity of this assumption is to look at type 3 tasks (that is, the tasks in which all options carried 10 points) and to measure how far subjects who saw the same labels gave matched responses. For the Amsterdam experiment, in which labels were randomised by session, we can calculate for each treatment an *average within-session NCI*. This is a weighted average of NCIs which have been calculated separately for each session. For the Nottingham experiment, in which labels were randomised separately for each pair of subjects, we count the number of cases in which the responses of paired subjects matched one another, and express this as a ratio of the expected number of matches under the assumption of random choice. If (consciously or unconsciously) subjects use labels as a means of matching, these measures will be greater than 1.

Table 4 presents these measures, averaged across all type 3 tasks, for the two experiments and the three treatments. It is clear that, for subjects who saw the same labels, matches were more frequent than would have been generated by random choice.[19] Surprisingly, in both experiments, pickers were more 'successful' in matching on the (intendedly) nondescript labels of the number tasks than on the apparently more distinguishable labels of the text tasks (compare the NCI_{PP} measures in table 2). However, in contrast to the text tasks, there is little evidence to suggest that the responses of guessers or coordinators are more concentrated than those of pickers. This seems to be a case in which matching, even among coordinators, is attributable to primary salience. In any event, these data suggest that if two partners both use labels as their method of trying to coordinate, their probability of success is only about 30 or 40 per cent greater than if they choose at random. Unless team-reasoning individuals have highly unrealistic expectations of their ability to discriminate between (what were intended as) nondescript labels, team-optimal choices in both type 1 and type 2 tasks will be as specified in section 2.4.

4.2 *Amsterdam results*

The results for type 1 tasks, shown in table 3, are extremely sharp. There are five such tasks. In every case, as one would expect, almost all pickers and guessers chose 10-point options. HL1 predicts that coordinators will choose 10-point options more frequently than if choices were made at random. In all cases, the opposite is true, contrary to cognitive hierarchy theory. HL2 identifies some other response as team-optimal and predicts that coordinators will choose this more frequently than if choices were made at random. In every case, more than 60 per cent of coordinators chose the team-optimal response, while if choices were random, the expected frequency would be only 0.167 or 0.25. In every case, the null hypothesis of random choice is rejected at the 1 per cent level. These results give very strong support to the team reasoning hypothesis.

In the five type 2 tasks, we again find that almost all pickers and guessers chose 10-point options. For these tasks, HL1 and HL2 make the same prediction, namely that coordinators will choose 10-point options more frequently than if choices were made at random. In three tasks (NA6, NA7 and NA8), overwhelming majorities of coordinators chose 10-point options, and the null hypothesis is rejected at the 1 per cent level. However, this prediction is not as successful for the other two tasks (NA9 and NA10). In these latter tasks, although 10-point options were chosen by large majorities of subjects (73 and 66 per cent respectively), the frequency of such choices was *less* than if subjects had acted at random.

Notice that, in each of the five type 2 tasks, there are just two numbers of points, 'high' (10) and 'low' (9 or 1, depending on the task). In NA6, NA7 and NA8, there are exactly as many high-point options as there are low-point ones; thus, apart from any differences in the labels themselves, the *only* distinguishing feature of the low-point options is that they carry fewer points. In NA9 and NA10, in contrast, there are more high-point options than low-point ones, and so a low-point option can be perceived as an odd one out. We speculate that a minority of subjects favoured the odd one out, even in cases in which this was not team-optimal.

4.3 *Nottingham results*

In the Nottingham experiment, as in the Amsterdam one, almost all pickers and guessers chose 10-point options in both type 1 and type 2 tasks. However, the responses that are relevant for our tests are those for coordinators; and here the Nottingham results are rather different.

In all six type 1 tasks, coordinators chose 10-point options slightly less frequently (and, correspondingly, chose the team-optimal 9-point option slightly more frequently) than if choices had been made at random; but, in each case, the null hypothesis of random choice cannot be rejected. Summing over all type 1 tasks, the 9-point option accounted for 23.9 per cent of choices, compared with the random-choice benchmark of 20 per cent. Formally, neither HL1 nor HL2 is supported by the evidence. However, an examination of behaviour at the level of the individual subject shows that coordinators' choices were *not* random. Of the 63 instances of subjects choosing 9-point options, 46 were attributable to just 8 of the 44 coordinators, while 25 coordinators never chose any 9-point option. We suggest that the most credible interpretation is that, on type 1 tasks, the majority of Nottingham subjects behaved roughly in accordance with cognitive hierarchy theory, while a minority behaved in similarly rough accordance with the theory of team reasoning.

In every type 2 task, the frequency with which coordinators chose 10-point options was higher than the random-choice benchmark, as predicted by both HL1 and HL2. The null hypothesis of random choice can be rejected at the 5 per cent level in three cases out of six. An analysis of behaviour at the individual level shows that the instances in which 1-point options were chosen were not random errors. All 25 of these instances were attributable to just 8 subjects, all of whom also chose at least one 9-point option in a type 1 task. The implication is that a small minority of subjects may have been attracted to odd-one-out options.

5 Taking stock

As an aide-memoire, the results from the two experiments are summarised in table 5. The Amsterdam results, for both text and number tasks, support the theory of team reasoning rather than cognitive hierarchy theory. The Nottingham results, for both text and number tasks, seem to point in the opposite direction.

Given the similarities of design between the two experiments, the differences between their results are surprising. In this and the following section, we consider possible explanations for these differences. In thinking about these explanations, it is important to keep in mind that the two experiments support independent tests of well-defined hypotheses derived from two recognised theories. The validity of those tests is unaffected by the analysis which follows. However, an investigation of the differences between the two sets of results may give some clues for further theoretical work.

One possibility is that there was a difference between the two subject pools with respect to their modes of reasoning about coordination and Hi-Lo games *in general*. Without completely ruling out this explanation, we judged it unlikely. It seemed implausible to suppose that the modes of reasoning used by Dutch university students are fundamentally different from those used by their English counterparts.

Another possibility is that differences between the displays may be having some effect. In the number tasks, the labels used in the Nottingham display (randomly generated patterns) may seem less nondescript than those of the Amsterdam display (positions of moving discs). This may have prompted Nottingham subjects to focus on labels rather than payoffs. But the normalised frequencies of matching in type 3 number tasks suggest that, in fact, the Nottingham labelling was only slightly less nondescript than the Amsterdam labelling (see section 4.1 and Table 4). A further difference is that the Nottingham display allows subjects to label objects by their positions in the row, using concepts such as 'first', 'middle', and 'last'. Because this layout is randomised independently for each co-player, position cannot be used as a coordinating device; but it might still influence behaviour in the picking treatment. To the extent that picking is influenced by position (which pickers' co-players cannot observe), successful guessing is made more difficult, as is coordination if players act according to cognitive hierarchy theory. But in fact, pickers' choices were distributed almost uniformly among the five positions; and Nottingham responses were *more* consistent with cognitive hierarchy theory than were Amsterdam responses.

We judged it most likely that the difference in results reflected some difference(s) in the content of the labels used in the two designs. Looking for clues in subjects' responses to text tasks, we tried to find common features in those labels that were the modal choices of coordinators. For the Amsterdam coordinators, the most obvious common feature seemed to be that of *standing out* from the other labels by virtue of some significant but not necessarily desirable characteristic. For the Nottingham coordinators, choices seemed to be concentrated on the labels whose content was most liked by, or was the *favourite* of, most students. But these interpretations were merely conjectures, based on our intuitions about the cultural significance of different labels. As we have said repeatedly, our investigative strategy is to avoid appeals to intuitions about salience, and instead to use cross-treatment comparisons in which cultural variables are held constant. To test our conjectures, we used two further treatments in a questionnaire study, which we now describe.

6 The questionnaire study

We administered a questionnaire to independent samples of respondents recruited from students at the Universities of Amsterdam and Nottingham. The aim was to investigate respondents' perceptions of the labels used in the original experiment in relation to the criteria of standing out and favouriteness. The study was carried out in 2002; respondents were rewarded by being entered in a lottery with a cash prize.

The questionnaire had two variants, which were administered to different, randomised samples. One variant investigated perceptions of standing out, the other perceptions of favouriteness. Each questionnaire contained a mixture of Amsterdam text tasks, Nottingham text tasks and Nottingham number tasks in randomised order. (Because the questionnaire was a pen-and-paper exercise, the computerised displays of the Amsterdam experiment could not be replicated.) For each task, the questionnaire displayed a row of four or five 'items'; these were the labels used in the relevant task in the original experiments (i.e. the strings of text which appeared on the 'discs' or in the upper 'boxes' of the text task displays, or the patterns in the upper boxes in the Nottingham number tasks; points were not shown). Respondents were given the following instructions (text which differed between the two variants of the questionnaire is shown in square brackets):

> In each task, you are shown a row of four or five 'items'. In each task, you must do one of two things.
>
> First, show [*which of the items is your favourite/ for you, which of the items stands out from the others*]. You show this by circling one of the items. You *must* circle exactly one item in each task. Even if you do not feel strongly that any of the items [is your favourite/ stands out], please circle one of them; the second part of the task will allow you to tell us how strong your feelings are.
>
> Second, show *how strongly you feel that the item you have circled* [*is your favourite/ stands out from the others*]. You show this by marking a point on a scale from 0 ('not at all') to 5 ('very strongly').

By using questionnaires with different sets of tasks, we were able to collect around 95 responses for each of the 28 text tasks in the original experiments, and for six representative examples of Nottingham number tasks. For each task, whichever experiment it was taken from, we collected approximately equal numbers of questionnaire responses in Amsterdam and Nottingham. This allowed us to check for subject pool effects. In fact, we found no systematic differences between responses collected in the two locations.[20] We therefore combined the two sets of responses.

Table 6 reports respondents' average strengths of feeling about standing-outness and favouriteness for the Amsterdam text tasks, Nottingham text tasks, and Nottingham number tasks. The most striking feature of these data is the lower strength of feeling for number tasks than for text tasks. This is consistent with our intention that the labels for number tasks be nondescript. A further feature is that standing-outness was more pronounced than favouriteness in the Amsterdam tasks, while the opposite was true of the Nottingham tasks. Hypotheses of no difference between reported strength of feeling between the two sets of text tasks are rejected at the 5 per cent level (in two-tailed Wilcoxon signed rank tests, $p = 0.026$ for standing-outness and $p < 0.01$ for favouriteness).

Using questionnaire responses to questions about standing out (S) and favourites (F) in conjunction with the actual responses of pickers (P), guessers (G) and coordinators (C) in the experiments, we can calculate within-group NCIs for S-responses (denoted NCI_{SS}) and F-responses (NCI_{FF}), and a cross-group NCI for each pair of distinct response groups. These statistics are shown in table 2.

Two features of these data, specific to text tasks, are of particular interest. First, consider NCI_{SS} and NCI_{FF}. For the Amsterdam tasks, the two indices have similar values. The value of NCI_{SS} for Nottingham tasks (1.398) is similar to the corresponding Amsterdam value, but NCI_{FF} (1.713) is much greater. The implication is that two co-players, trying to coordinate on a Nottingham task, would be much more likely to succeed by using the primary-salience rule ‘Choose your favourite’ than ‘Choose the label which stands out most for you’. Presumably they would be even more likely to succeed by using secondary salience (‘Choose the object you believe to be most people’s favourite’). It seems that, for the Nottingham tasks, choosing according to secondary salience may also be the best rule in Schelling’s sense: cognitive hierarchy theory and team reasoning make the same predictions about coordination.

Now consider the values of NCI_{CS} and NCI_{CF}. For the Amsterdam tasks, $\text{NCI}_{CS} = 1.521$ and $\text{NCI}_{CF} = 1.072$. The implication is that, on these tasks, coordinators tended to choose labels that were generally perceived as standing out, rather than ones that were generally perceived as favourites. For the Nottingham tasks, the opposite is true: $\text{NCI}_{CS} = 1.565$ and $\text{NCI}_{CF} = 1.838$.

To test this account of the behaviour of coordinators more rigorously, we estimated (separately for each experiment) an equation in which the dependent variable is the frequency with which each label was chosen in the coordination treatment, and the independent variables are the frequencies with which the same label was named as the stander-out (‘standout’) and as the favourite (‘favourite’). Since the frequency with which the *n*th option is chosen is a residual, where *n* is the number of options in a task, we dropped one observation for each task. Since the dependent variable is a proportion, we use a GLM specification with logistic link function, binomial error structure and robust standard errors (Papke and Wooldridge, 1996). Results are given in table 7. These regressions confirm that coordinators were attracted to standing-out labels in the Amsterdam experiment but to favourite labels in the Nottingham experiment.

In the light of the questionnaire data, we conjecture that the crucial difference between the two experiments is to be found in the specifications of the labels for the text tasks. In the Amsterdam tasks, the odd-one-out labels were perceived as standing out from the others. These labels were used as focal points, contrary to the predictions of cognitive hierarchy theory. The Nottingham tasks did not have such obvious odd-ones-out, while being more effective than the Amsterdam ones in priming ideas of relative desirability and favouriteness.[21] As a result (and consistent with both theories), favourites were focal points. We conjecture that there was some tendency for the modes of reasoning used in the text tasks to 'spill over' to the number tasks. In the Amsterdam experiment, coordinators consistently chose team-optimal options in number tasks, whether or not those options carried the highest number of points. In the Nottingham experiment, the majority of coordinators responded to number tasks by using the same favourite-based reasoning as they used in text tasks, with only a small minority choosing according to team optimality.

7 Discussion

The main aim of the two experiments was to test cognitive hierarchy theory and the theory of team reasoning as rival explanations of behaviour in pure coordination and Hi-Lo games. Formally, our conclusion must be that each theory failed at least one test. Cognitive hierarchy theory was disconfirmed in the Amsterdam text and number tasks, while both theories were disconfirmed in the Nottingham number tasks. However, it would be equally true to say that each theory had some success. Whenever one of the theories failed, the disconfirming evidence revealed a regularity in behaviour for which the other theory provided an explanation.

Our experiments seem to have identified two modes of reasoning, each of which is sometimes used by players of coordination and Hi-Lo games. Which of these modes of reasoning is brought into play may be sensitive to the decision context. In thinking about the relationship between the two theories, it is useful to step back and compare their main properties.

In the context of coordination and Hi-Lo games, perhaps the most important feature of cognitive hierarchy theory is the role it gives to players' pre-reflective inclinations – the non-rational choice propensities that generate primary salience, and that are modelled in the behaviour of 'level 0' players. The workings of the theory are such that equilibrium selection is strongly influenced by these inclinations. In contrast, the theory of team reasoning gives no role to primary salience. It models the decisions of agents who are fully rational, but in the special sense that they optimise over profiles of strategies assessed from the viewpoint of the players as a collective, rather than over individual strategies assessed from the viewpoints of individual players.

Intuitively, it seems that each of these approaches captures a significant aspect of focal points. On the one hand, many apparently obvious focal points seem to be identified by pre-reflective inclinations. For example, think of the pure coordination game in which the set of labels is {«heads», «tails»}. Why do most players choose «heads»? Even if we assume it to be common cultural knowledge that «heads» takes priority over «tails», this does not provide a *reason* for the players (individually or collectively) to choose «heads». Ultimately, it seems, an explanation has to appeal to a pre-reflective association between the idea of priority and the idea of choosing: it is psychologically more natural to choose the more important than to choose the less. On the other hand, there are equally obvious cases in which focal points seem to be identified by optimising over strategy profiles. Consider a pure coordination game in which each player has to point to one of four cubes on a tray; three are red and one is green. What makes the choice of the green cube the focal point? Here, it seems implausible to appeal to a pre-reflective propensity to pick the odd one out. The most obvious answer is that 'Choose the green cube' is a better rule for the two players together than 'Pick a red cube'.

Our experimental strategy was to create games in which these two ways of trying to identify focal points pull in different directions. In designing the text tasks, we tried to ensure that subjects' pre-reflective inclinations would attract them towards desirable or favourite labels, while thoughts about the best rule for the two co-players together would attract them to less desirable odd ones out. In the 'type 1' number tasks, the objects with the highest numbers of points are the most immediately desirable, but the best rule for the co-players together is to choose an object with fewer points. It seems that both of these forces were at work in our experiments, and that their relative strength was sensitive to details of experimental design. We seem to have found a class of coordination problems which, because they prime opposing modes of reasoning, are particularly difficult for people to solve.[22]

One of the most remarkable features of our experiments is the success with which participants overcame this difficulty. Within each of our experiments, a large majority of subjects used a common mode of reasoning for identifying focal points. That common mode of reasoning was different in the two experiments; but, as far we can tell, this difference was not attributable to differences between the subject pools. The implication is that our subjects were able to use subtle features of the experimental environment to solve the problem of coordinating *on a common mode of reasoning*. This behaviour reveals an ability to solve coordination problems at a conceptual level above that of the theories of cognitive hierarchy and team reasoning that we have been examining. Each of those theories captures certain aspects of focal-point reasoning, but some essential feature of the human ability to solve coordination problems seems to have escaped formalisation.

However disheartening this conclusion may be for game theorists, it ought not to be too surprising to readers of Schelling. In *Strategy of Conflict*, Schelling repeatedly insists on the diversity of the methods by which people find focal points, and rejects any suggestion that these methods can be reduced to a single formal theory. Although there are hints of team reasoning in his idea that players search for the 'best rule', he allows this search to range over a much wider domain than that represented in game theory (whether as practised in 1960, or as practised now). His list of methods or rules includes 'analogy', 'precedent', 'aesthetic or geometric configuration', 'casuistic reasoning', and 'whimsy' (1960, p. 57), and he even suggests that some methods use 'excuses' and 'pretences' in place of reasons and beliefs (p. 298). In the context of a Nash demand game, he says:

> The basic intellectual premise, or working hypothesis, for rational players in this game seems to be the premise that some rule must be used if success is to exceed coincidence, and that the best rule to be found, *whatever its rationalization*, is consequently a rational rule. (p. 283, italics added)

Schelling is advising us not to expect a unified theoretical rationalisation of focal points.[23] Nearly half a century later, we are beginning to understand some of the methods by which focal points are found, but so far the search for a unified theory has been unsuccessful. We suspect that Schelling is not surprised by this state of affairs.

Table 1: responses to text tasks

		pick	guess	coordinate
Amsterdam tasks				
TA1	grijs	11	17	35
	indigo	12	8	8
	karmozijn	9	9	5
	magenta	10	11	1
	turkoois	11	7	7
	NCI	0.935	1.040	2.125
	significance wrt: pick		ns	**
	guess			**
	chisq wrt guess			**
TA2	Ford	10	11	31
	Ferrari	13	22	11
	Jaguar	20	7	5
	Porsche	10	12	9
	NCI	1.040	1.124	1.472
	significance wrt: pick		ns	**
	guess			*
	chisq wrt guess			**
TA3	Mannheim	12	9	25
	Berlin	3	2	4
	Brussel	8	13	9
	Lissabon	15	8	7
	Madrid	15	20	11
	NCI	1.115	1.255	1.355
	significance wrt: pick		ns	ns
	guess			ns
	chisq wrt guess			*
TA4	peanut	13	13	12
	almond	13	20	18
	cashew	20	14	15
	walnut	7	5	11
	NCI	1.064	1.112	0.984
	significance wrt: pick		ns	#
	guess			#
	chisq wrt guess			ns
TA5	glass	11	14	30
	diamond	24	28	21
	emerald	7	8	3
	sapphire	11	2	2
	NCI	1.180	1.504	1.684
	significance wrt: pick		ns	*
	guess			ns

	chisq wrt guess			*
TA6	plastic	15	16	36
	chrome	11	11	7
	copper	9	10	2
	iron	11	6	6
	steel	7	9	5
	NCI	0.985	1.020	2.200
	significance wrt: pick		ns	**
	guess			**
	chisq wrt guess			**
TA7	bread	8	6	8
	curry	12	9	23
	pizza	10	16	17
	steak	23	21	8
	NCI	1.136	1.148	1.156
	significance wrt: pick		ns	ns
	guess			ns
	chisq wrt guess			**
TA8	water	20	15	38
	beer	13	26	11
	sherry	4	1	0
	whisky	6	6	5
	wine	10	4	2
	NCI	1.210	1.700	2.495
	significance wrt: pick		*	**
	guess			**
	chisq wrt guess			**
TA9	Carlsberg	25	23	37
	Corsendonk	5	3	2
	Grimbergen	8	13	2
	Rochefort	9	4	11
	Westmalle	6	9	4
	NCI	1.410	1.420	2.365
	significance wrt: pick		ns	**
	guess			**
	chisq wrt guess			**
TA10	frog	17	17	41
	leopard	11	11	5
	panther	7	4	5
	tiger	18	20	5
	NCI	1.060	1.168	2.208
	significance wrt: pick		ns	**
	guess			**
	chisq wrt guess			**

TA11	bicycle	18	18	37
	aeroplane	19	18	15
	helicopter	6	6	2
	hovercraft	10	10	2
	NCI	1.116	1.104	2.008
	significance wrt: pick		#	**
	guess			**
	chisq wrt guess			**
TA12	chess	18	15	36
	football	11	30	14
	squash	5	1	0
	tennis	16	6	3
	volleyball	3	0	3
	NCI	1.235	2.095	2.360
	significance wrt: pick		**	**
	guess			ns
	chisq wrt guess			**
TA13	Bern	11	12	29
	Barbados	13	10	4
	Florida	21	17	12
	Honolulu	8	13	11
	NCI	1.076	0.980	1.384
	significance wrt: pick		#	*
	guess			**
	chisq wrt guess			*
TA14	sitting	16	21	39
	jogging	6	5	5
	running	20	15	10
	walking	11	11	2
	NCI	1.104	1.148	2.072
	significance wrt: pick		ns	**
	guess			**
	chisq wrt guess			**

Nottingham tasks

TN1	Friday lunchtime	13	2	4
	Monday morning	6	2	3
	Saturday night	17	36	34
	Sunday night	4	1	0
	Wednesday evening	5	4	3
	NCI	1.235	3.220	3.030
	significance wrt: pick		**	**
	guess			#
	chisq wrt guess			ns
TN2	Earth	18	25	33

	Mars	5	6	3
	Mercury	9	2	1
	Saturn	8	4	3
	Venus	5	8	4
	NCI	1.195	1.770	2.855
	significance wrt: pick		*	**
	guess			**
	chisq wrt guess			ns
TN3	Ford	3	4	2
	Mercedes	8	11	13
	Pontiac	4	1	0
	Porsche	21	29	26
	Volkswagen	9	0	3
	NCI	1.430	2.360	2.150
	significance wrt: pick		**	*
	guess			#
	chisq wrt guess			ns
TN4	plain omelette	5	9	7
	cheese omelette	19	19	21
	ham omelette	6	11	3
	mushroom omelette	8	2	7
	prawn omelette	7	4	6
	NCI	1.235	1.360	1.425
	significance wrt: pick		ns	ns
	guess			ns
	chisq wrt guess			ns
TN5	1	4	3	5
	2	5	0	2
	7	9	16	6
	10	11	7	8
	15	16	19	23
	NCI	1.145	1.590	1.625
	significance wrt: pick		*	*
	guess			ns
	chisq wrt guess			ns
TN6	stool	4	3	5
	deck chair	11	8	7
	dining chair	4	2	1
	easy chair	10	24	15
	rocking chair	16	8	16
	NCI	1.170	1.695	1.355
	significance wrt: pick		*	ns
	guess			#
	chisq wrt guess			ns
TN7	Ontario	9	3	4

	Colorado	8	6	5
	Florida	18	33	33
	Louisiana	4	0	0
	Nevada	6	3	2
	NCI	1.200	2.775	2.880
	significance wrt: pick		**	**
	guess			ns
	chisq wrt guess			ns
TN8	sitting	3	2	2
	jogging	11	4	4
	sunbathing	13	20	26
	swimming	11	5	7
	walking	7	4	5
	NCI	1.070	2.315	1.920
	significance wrt: pick		**	**
	guess			#
	chisq wrt guess			ns
TN9	2000	13	23	27
	1978	3	1	2
	1979	3	1	2
	1980	11	7	5
	1981	15	13	8
	NCI	1.230	1.780	2.065
	significance wrt: pick		**	**
	guess			ns
	chisq wrt guess			ns
TN10	John	9	10	14
	David	11	8	9
	Michael	9	13	12
	Robert	9	6	5
	Steven	7	8	4
	NCI	0.930	0.980	1.105
	significance wrt: pick		ns	ns
	guess			ns
	chisq wrt guess			ns
TN11	win nothing	4	2	1
	win champagne	8	4	1
	win chocolate	6	1	0
	win money	22	38	41
	win trophy	5	0	1
	NCI	1.465	3.585	4.335
	significance wrt: pick		**	**
	guess			ns
	chisq wrt guess			ns
TN12	red	9	15	8

	blue	17	16	23
	green	7	3	4
	orange	3	1	3
	purple	9	10	6
	NCI	1.170	1.380	1.610
	significance wrt: pick		ns	*
	guess			ns
	chisq wrt guess			ns
TN13	carrot juice	5	6	2
	apple juice	10	19	29
	grapefruit juice	7	4	3
	mango juice	12	7	5
	pineapple juice	11	9	5
	NCI	0.995	1.260	2.275
	significance wrt: pick		ns	**
	guess			**
	chisq wrt guess			ns
TN14	Calais	5	1	4
	Berlin	2	3	1
	Paris	10	23	27
	Prague	8	7	2
	Rome	20	11	10
	NCI	1.385	1.675	2.130
	significance wrt: pick		ns	**
	guess			ns
	chisq wrt guess			ns

Table 2: normalised coordination indices

Amsterdam text tasks

	pick	guess	coordinate	stand out	favourite
pick	1.116	1.172	1.191	1.164	1.165
guess		1.259	1.267	1.227	1.202
coordinate			1.819	1.521	1.072
stand out				1.334	1.181
favourite					1.397

Nottingham text tasks

	pick	guess	coordinate	stand out	favourite
pick	1.204	1.416	1.470	1.187	1.345
guess		1.982	2.067	1.492	1.756
coordinate			2.197	1.565	1.838
stand out				1.398	1.438
favourite					1.713

Nottingham number tasks

	stand out	favourite
stand out	1.274	1.073
favourite		1.095

Table 3: responses to type 1 and type 2 number tasks

		pick	guess	coordinate	significance
Type 1 tasks: Amsterdam					
NA1	10x3	50	48	8	#
	9x1	3	4	48	**
NA2	10x5	49	46	10	#
	9x1	4	6	46	**
NA3	10x3	51	50	10	#
	9x1	1	1	43	**
	8x1	0	0	1	
	7x1	1	1	2	
NA4	10x3	52	48	15	#
	9x2	0	1	0	
	8x1	1	3	41	**
NA5	10x4	49	51	21	#
	9x2	4	1	35	**
Type 1 tasks: Nottingham					
NN1	10x4	42	44	35	#
	9x1	3	1	9	ns
NN2	10x4	44	44	33	#
	9x1	1	1	11	ns
NN3	10x4	43	44	34	#
	9x1	2	1	10	ns
NN4	10x4	43	43	34	#
	9x1	2	2	10	ns
NN5	10x4	44	40	33	#
	9x1	1	5	11	ns
NN6	10x4	44	42	32	#
	9x1	1	3	12	ns

Type 2 tasks: Amsterdam

NA6	10x1	52	51	54	**
	9x1	1	1	2	
NA7	10x3	52	49	50	**
	9x3	1	3	6	
NA8	10x1	52	51	54	**
	1x1	1	1	2	
NA9	10x3	49	51	41	#
	1x1	4	1	15	
NA10	10x5	47	51	37	#
	1x1	6	1	19	

Type 2 tasks: Nottingham

NN7	10x4	44	45	41	*
	1x1	1	0	3	
NN8	10x4	44	44	40	*
	1x1	1	1	4	
NN9	10x4	41	45	42	**
	1x1	4	0	2	
NN10	10x4	44	44	38	#
	1x1	1	1	6	
NN11	10x4	45	43	39	#
	1x1	0	2	5	
NN12	10x4	45	43	39	#
	1x1	0	2	5	

Table 4: normalised frequency of matching on type 3 number tasks

	pick	guess	coordinate
Amsterdam	1.242	1.286	1.336
Nottingham	1.364	1.515	1.402

Table 5: summary of results

	cognitive hierarchy theory predicts	team reasoning theory predicts	Amsterdam result	Nottingham result
tests using pure coordination games:				
compare NCI_{GG} and NCI_{PP} (test PC1)	$NCI_{GG} > NCI_{PP}$	$NCI_{GG} > NCI_{PP}$	$NCI_{GG} > NCI_{PP}$ (difference small)	$NCI_{GG} > NCI_{PP}$ (difference large)
compare distributions of responses of coordinators and guessers (test PC2)	coordinators' and guessers' responses have same distribution	no prediction	distributions different	no significant differences between distributions
if distributions different, compare NCI_{CC} and NCI_{GG} (test PC3)	not applicable	$NCI_{CC} > NCI_{GG}$	$NCI_{CC} > NCI_{GG}$	not applicable
tests using nondescript Hi-Lo games:				
type 1 tasks (high-payoff and team-optimal labels are different: tests HL1 and HL2)	high-payoff labels chosen with greater-than-random frequency	team-optimal labels chosen with greater-than-random frequency	team-optimal labels chosen with greater-than-random frequency	both types of label chosen with approximately random frequency
type 2 tasks (high-payoff labels are also team-optimal: tests HL1 and HL2)	high-payoff labels chosen with greater-than-random frequency	high-payoff labels chosen with greater-than-random frequency	high-payoff labels usually chosen with greater-than-random frequency	high-payoff labels chosen with greater-than-random frequency

Table 6: strength of standing-outness and favouriteness (on 0-5 scale)

	standing-outness	favouriteness
Amsterdam text tasks	3.45	3.39
Nottingham text tasks	3.31	3.54
Nottingham number tasks	2.23	1.92

Table 7: proportion of coordinators choosing an option regressed on its favouriteness and standing-out score

	Coefficient (SE)	**∂y/∂x (SE)**
Amsterdam experiment		
standout	0.089** (0.005)	0.014** (0.001)
favourite	-0.026** (0.003)	-0.004** (0.001)
constant	-2.829** (0.179)	

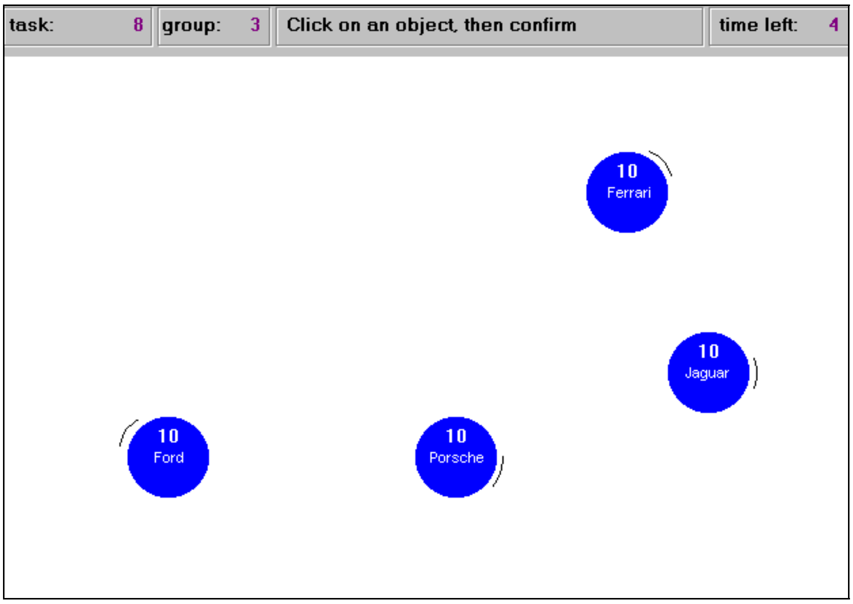
LogPseudoL: -15.06; AIC: 0.75; BIC -13.27

Nottingham experiment		
standout	0.018 (0.017)	0.002 (0.002)
favourite	0.057** (0.012)	0.008** (0.002)
constant	-3.092** (0.200)	

LogPseudoL: -15.98; AIC: 0.68; BIC: -4.45

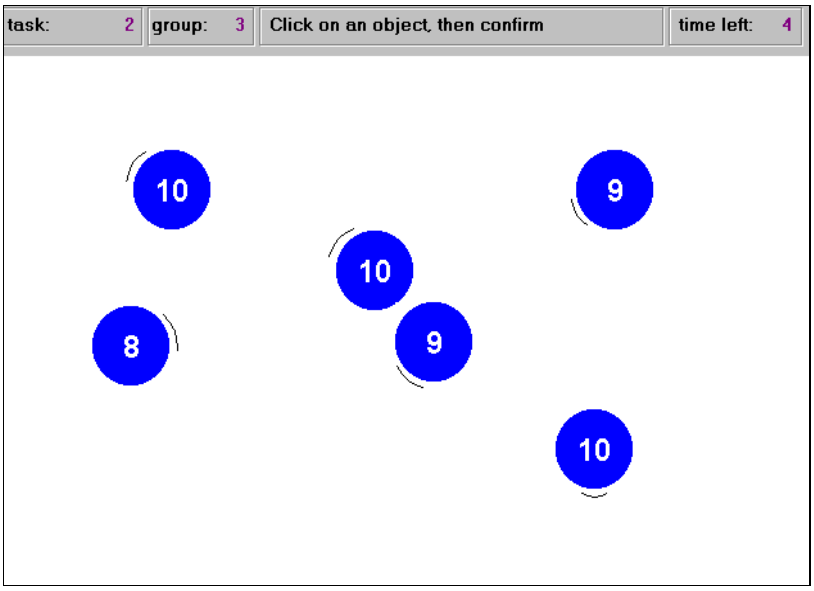
Notes 1. ** denotes significance at the 1% level.
2. ∂y/∂x calculated at mean values of the independent variables.

Figure 1: display for Amsterdam text tasks



task: 8 group: 3 Click on an object, then confirm time left: 4

Figure 2: display for Amsterdam number tasks



task: 2 group: 3 Click on an object, then confirm time left: 4

Figure 3: display for Nottingham text tasks

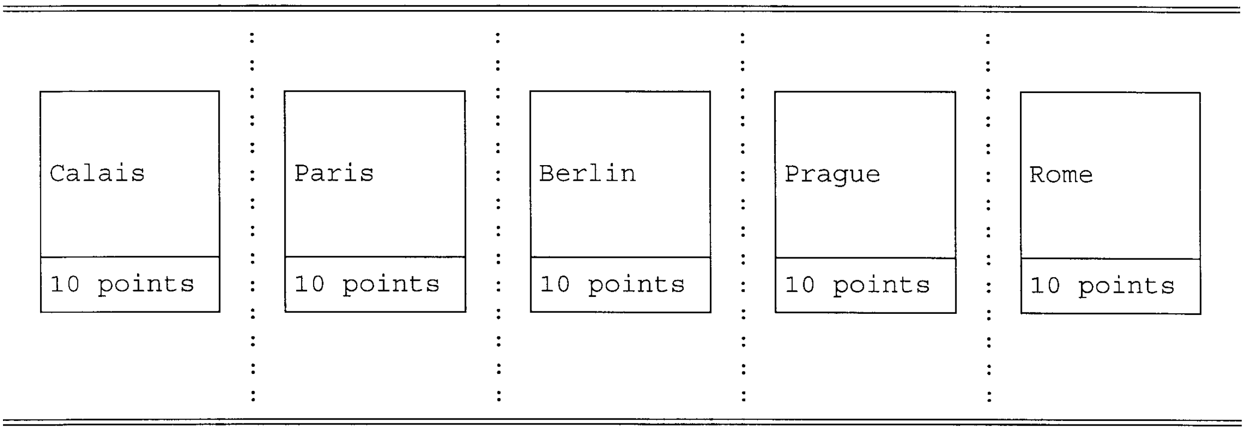
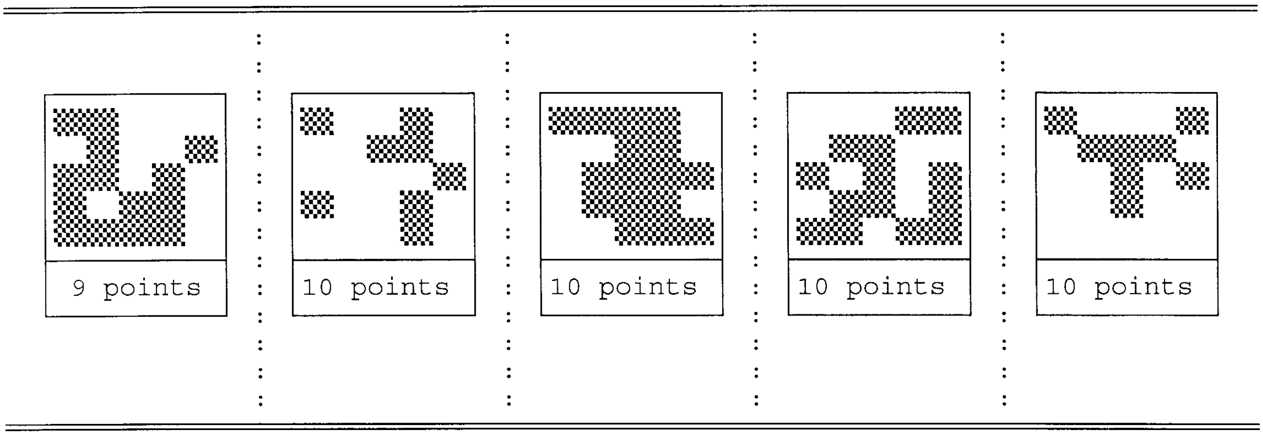


Figure 4: display for Nottingham number tasks



References

Bacharach, Michael (1993). Variable universe games. In Ken Binmore et al. (eds.) *Frontiers of Game Theory*. Cambridge, MA: MIT Press.

Bacharach, Michael (1999). Interactive team reasoning: a contribution to the theory of cooperation. *Research in Economics* 53: 117–147.

Bacharach, Michael (2006). *Beyond Individual Choice: Teams and Frames in Game Theory*. Edited by Natalie Gold and Robert Sugden. Princeton University Press.

Bacharach, Michael and Michele Bernasconi (1997). The variable frame theory of focal points: an experimental study. *Games and Economic Behavior* 19: 1-45.

Bacharach, Michael and Dale Stahl (2000). Variable-frame level-*n* theory. *Games and Economic Behavior* 33: 220-246.

Binmore, Ken and Larry Samuelson (2006). The evolution of focal points. *Games and Economic Behavior* 55: 21-42.

Camerer, Colin F., Teck Ho and Kuan Chong (2004). A cognitive hierarchy model of games. *Quarterly Journal of Economics* 119: 861-898.

Casajus, André (2001). *Focal Points in Framed Games: Breaking the Symmetry*. Berlin: Springer-Verlag.

Crawford, Vincent and Hans Haller (1990). Learning how to cooperate: optimal play in repeated coordination games. *Econometrica* 58: 571-595.

Crawford, Vincent and Nagore Iriberri (2007). Fatal attraction: salience, naïveté, and sophistication in experimental "hide-and-seek" games. Working paper, University of California, San Diego.

Cubitt, Robin P. and Sugden, Robert (2003). Common knowledge, salience and convention: a reconstruction of David Lewis's game theory. *Economics and Philosophy* 19: 175–210.

Efron, Bradley (1979). Bootstrap methods: another look at the jacknife. *Annals of Statistics* 7: 1–26.

Crawford, Vincent, Uri Gneezy and Yuval Rottenstreich (2007). The power of focal points is limited: even minute payoff asymmetry may yield large coordination failures. University of Chicago Business School.

Harsanyi, John and Reinhard Selten (1988). *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press.

Janssen, Maarten (2001). Rationalising focal points. *Theory and Decision* 50: 119-148.

Lewis, David (1969). *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Mehta, Judith, Chris Starmer and Robert Sugden (1994). The nature of salience: an experimental investigation. *American Economic Review* 84: 658–673.

Papke, Leslie and Jeffrey Wooldridge (1996). Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics* 11: 619–632.

Schelling, Thomas (1960). *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Stahl, Dale O. and Paul Wilson (1995). On players' models of other players. *Games and Economic Behavior* 10: 218–254.

Sugden, Robert (1993). Thinking as a team: towards an explanation of non-selfish behavior. *Social Philosophy and Policy* 10: 69–89.

Sugden, Robert (1995). A theory of focal points. *Economic Journal* 105: 533–550.

Sugden, Robert and Zamarrón, Ignacio (2006). Finding the key: the riddle of focal points. *Journal of Economic Psychology* 27: 609-621.

Notes

[1] From now on, we will use the term ‘cognitive hierarchy theory’ to refer to this approach in general, and not merely to the specific model proposed by Camerer et al.

[2] Repeated coordination games provide additional, confounding means of communication, which are not part of our subject matter. This class of games is analysed by Crawford and Haller (1990).

[3] This term is due to Bacharach (2006). Bacharach uses it only for cases in which there is some j such that $U_j > U_k$ for all $k \neq j$, but (as we shall show in section 1.4) his analysis can be extended to the wider class of games encompassed by our definition.

[4] Lewis presented his ideas relatively informally, at a time when some of what are now seen as fundamental principles of game theory had not been developed. Mehta et al (1994) discuss Lewis’s theory of salience. For a fuller discussion of Lewis’s game theory, see Cubitt and Sugden (2003).

[5] In an alternative formulation, proposed by Crawford and Iriberri (2007), a level 2 player believes that her opponent reasons at level 1, a level 3 player believes that his opponent reasons at level 2, and so on. For the games analysed in this paper, the implications of the two versions of the theory are essentially the same.

[6] The two hypotheses have different implications when the profile of strategies that is best for the group is not a Nash equilibrium, as in the Prisoner’s Dilemma.

[7] Using this approach as their representation of focal-point reasoning, Binmore and Samuelson (2006) develop an evolutionary model in which ‘monitoring’ of labels is costly; selection induces an equilibrium in which the degree of monitoring is less than optimal.

[8] This concept of ‘nondescriptness’ is due to Bacharach and Bernasconi (1997). An alternative implementation of Schelling’s idea, discussed by Crawford and Haller (1990), is to assign labels to the two co-players by independent random draws from a given distribution. We take the view that, if the Crawford–Haller labelling system is used, Schelling’s game is no longer the 4x4 Hi-Lo game that the payoff matrix purports to represent. Instead, there is a 4x4 payoff matrix in which one cell has the payoff profile (9, 9) and nine cells have the payoff profile (10/3, 10/3). Since we are investigating classic coordination games, we use nondescript labelling rather the Crawford–Haller method.

[9] A different possibility is suggested by Crawford and Iriberri’s (2007) assumption that level 0 reasoners choose options that, in some intuitive sense, stand out. Conceivably, the fact that l_4 is an odd-one-out payoff might it *more* likely to be chosen by level 0 reasoners. In fact, our results do not support this version of cognitive hierarchy theory. (In relation to pure coordination games, the best evidence of behaviour in accordance with cognitive hierarchy theory comes from the Nottingham

experiment. In that experiment, in the nondescript Hi-Lo games most similar to the present example, the low-payoff label does *not* act as a focal point: see the results for tasks NN1 to NN6 in Table 3.)

[10] The Amsterdam experiment was carried out by Bardsley, the Nottingham one by Mehta, Starmer and Sugden. In the early stages of the development of the design, all four authors were working together in the UK. The design process bifurcated when Bardsley moved to the Netherlands. As a result, the two experiments have a common basic design, implemented in slightly different ways.

[11] When different objects carry different numbers of points, the picking treatment cannot be interpreted as eliciting $\boldsymbol{p}^0$. If cognitive hierarchy theory holds, reasoners of level 1 and above will not perceive tasks in the picking treatment as 'just picking': they will recognise the rationality of choosing an object with the maximum number of points.

[12] Labels were written in English whenever the relevant words would be familiar to Dutch students. Subjects were given a list of translations between English and Dutch for all labels. The relevant Dutch-to-English translations are: grijs = grey, karmozijn = crimson, turkoois = turquoise, Lissabon = Lisbon.

[13] Carlsberg is a popular and widely available brand of beer. Corsendonk, Grimbergen, Westmalle and Rochefort are specialist Trappist-style beers, brewed in Belgium.

[14] In the experiment of Mehta et al, most questions were open-ended (e.g. 'Name any car manufacturer'), rather than requiring a closed choice from a finite set of pre-specified labels. With the benefit of hindsight, we now think that the two types of question may prompt different rules of selection. For example, Ford (the archetypal car manufacturer) was the clear focal point for Mehta et al's open-ended task; but, facing a finite list of car manufacturers, subjects may perceive 'Choose the most glamorous' as a more obvious rule.

[15] The intended salients for TA1, …, TA14 were: «grijs», «Ford», «Mannheim», «peanut», «glass», «plastic», «bread», «water», «Carlsberg», «frog», «bicycle», «chess», «Bern», and «sitting» (all odd ones out). For TN2, …, TN14 they were: «Earth» (status quo), «Ford» (archetype), «plain omelette» (archetype), «1» (smallest), «stool» (odd one out), «Ontario» (odd one out), «sitting» (status quo), «2000» (round number, most talked about), «John» (archetype), «win nothing» (odd one out), «red» (archetype), «carrot juice» (odd one out), and «Calais» (odd one out). TN1 did not have an intended salient. The booklets were originally prepared for an experiment carried out in 2000 at the University of East Anglia on a *Friday lunchtime*, making this the status quo in TN1. In that experiment, subjects in the *picking* treatment of the number tasks distributed their choices approximately randomly between options, irrespective of the points assigned to them. Since the most credible explanation of this behaviour was that subjects had not understood the role of points in the experiment, we revised the instructions and re-ran the experiment in Nottingham, using the same

booklets. The Nottingham experiment took place on a Wednesday lunchtime. The results of the East Anglia experiment are available from the authors on request.

[16] Our claims about team reasoning require the assumptions (i) that the safer of two lotteries is preferred when its expected value is greater than that of the riskier lottery and (ii) that the riskier lottery is preferred when its expected value is at least twice that of the safer one.

[17] If, for a given label, the expected number of choices is less than 5, we combine choices for that label with those for the label with the next smallest expected number of choices, and so on until the expected number of entries in each cell is at least 5.

[18] This 'average NCI' is calculated as follows. For each task independently, we calculate the probability that two pickers, drawn at random without replacement, choose the same option. We sum these probabilities across all tasks to arrive at the expected number of 'same choices' in the whole experiment, per pair of pickers. We then divide this by the expected number of 'same choices' if pickers choose at random.

[19] The method of averaging is the same as that used to generate the summary statistics shown in Table 2. For the Amsterdam experiment, we carried out a bootstrap test, separately for each of the four type 3 tasks and for each treatment, of whether the average within-session NCI was greater than 1. A statistically significant difference (always at the 1 per cent level) was found for two tasks in the picking treatment, two tasks in the guessing treatment, and three tasks in the coordination treatment. In the Nottingham experiment, the number of pairs of coordinators (22) and the number of matching responses (on average, about 6 per task) is too small for powerful statistical tests at the pair or task level. Aggregating over the 22 pairs and the 6 tasks (i.e. 132 cases), there were 37 matched responses, compared with an expectation of 26.4 from random choice. The null hypothesis that all coordinators chose randomly can be rejected at the 5 per cent level in a binomial test.

[20] Statistically significant differences (at the 5 per cent level, using a chi-squared test) between responses in Amsterdam and Nottingham were found for only 3 of the 34 standing-out questions and for only 5 of the 34 questions about favourites.

[21] This effect may have been enhanced by the fact that the Nottingham instructions used the word 'choose' to refer to the act of selecting a label, while the Amsterdam instructions used the more neutral expression 'click on'. *Choosing* has stronger connotations of liking (and hence of favouriteness) than *clicking on*.

[22] There is perhaps some parallel here with the results of an experiment reported by Crawford, Gneezy and Rottenstreich (2007), in which small asymmetries in payoffs between players disrupt focal-point reasoning. In that experiment, however, there seems to be a motivational conflict between seeking to coordinate with one's co-player and seeking advantage (or avoiding

disadvantage) relative to her. In our experiments, the conflict is between opposing modes of reasoning about how to coordinate.

[23] This reading of Schelling is defended by Sugden and Zamarrón (2006).