

# A POPPERIAN TEST OF LEVEL-K THEORY

By Shaun Hargreaves Heap\*, David Rojo  
Arjona\*\* and Robert Sugden\*

\* School of Economics, University of East Anglia

\*\*Economic Science Institute, Chapman University

## Abstract

We report an experimental test of level-k theory, applied to three simple games with non-neutral frames – Coordination, Discoordination and Hide and Seek. Using the same frame for all three games, we derive hypotheses that apply across the games and are independent of prior assumptions about salience. Those hypotheses are not confirmed by our experimental results. Our findings contrast with previous research which has fitted parameterised level-k models to Hide and Seek data. We show that, as a theory-testing criterion, the existence of a plausible model that replicates the main patterns in these data has a high probability of false positives.

## JEL classification codes

C72; C78; C91

## Keywords

level-k theory; Popper; Hide and Seek; coordination; discoordination

# A POPPERIAN TEST OF LEVEL-K THEORY\*

SHAUN HARGREAVES HEAP

DAVID ROJO ARJONA

ROBERT SUGDEN

Corresponding author: Robert SUGDEN, School of Economics, University of East Anglia, Norwich NR4 7TJ, United Kingdom; [r.sugden@uea.ac.uk](mailto:r.sugden@uea.ac.uk)

Shaun HARGREAVES HEAP, School of Economics, University of East Anglia, Norwich NR4 7TU, United Kingdom; [s.hargreavesheap@uea.ac.uk](mailto:s.hargreavesheap@uea.ac.uk)

David ROJO ARJONA, Economic Science Institute, Chapman University, One University Drive, Orange, CA 92866, United States [rojoarjo@chapman.edu](mailto:rojoarjo@chapman.edu)

*Abstract:* We report an experimental test of level- $k$  theory, applied to three simple games with non-neutral frames – Coordination, Discoordination and Hide and Seek. Using the same frame for all three games, we derive hypotheses that apply across the games and are independent of prior assumptions about salience. Those hypotheses are *not* confirmed by our experimental results. Our findings contrast with previous research which has fitted parameterised level- $k$  models to Hide and Seek data. We show that, as a theory-testing criterion, the existence of a plausible model that replicates the main patterns in these data has a high probability of false positives.

*Keywords:* level- $k$  theory; Popper; Hide and Seek; coordination; discoordination

*JEL classifications:* C72 (noncooperative games); C78 (bargaining theory, matching theory); C91 (design of experiments: laboratory, group behaviour)

*Date:* 26 September 2012

*Word count:* 12,507

\* We thank Vincent Crawford, Nagore Iriberri, Ariel Rubinstein, and many seminar and conference participants for comments on earlier versions of this paper.

## A POPPERIAN TEST OF LEVEL- $k$ THEORY

Traditionally, game theory has analysed the interaction of ideally rational agents whose rationality is common knowledge. Recently, however, there has been a growth of interest in investigating how *in fact* human agents reason in strategic situations, and how far this reasoning supports the solution concepts proposed by classical game theory. One of the leading theories of boundedly rational strategic reasoning is *level- $k$  theory* (Dale Stahl and Paul Wilson, 1994; Rosemarie Nagel, 1995; Miguel Costa-Gomes, Vincent Crawford, and Bruno Broseta, 2001). This theory distinguishes types of players according to the number of iterations of best-response reasoning they employ. ‘Level 0’ ( $L0$ ) players do not use best-response reasoning at all. The normal default assumption is that such players pick strategies at random, but in some models they are assumed to respond naïvely to non-strategic features of the relevant game. When games have non-neutral frames – that is, when strategies have labels with psychologically or culturally significant connotations – it is usually assumed that  $L0$  players favour the strategies with the most salient labels.  $L1$  players are assumed to know the probability distribution of  $L0$  choices and to choose best replies to this;  $L2$  players choose best replies to  $L1$  choices, and so on.<sup>1</sup> Thus, the aggregate behaviour of players is fully determined by the specification of  $L0$  and by the probability distribution over types. In this paper, we report an experimental test of level- $k$  theory as applied to a class of simple games with non-neutral frames.

Our methodological strategy is based on two Popperian principles: that a proposed explanation of any phenomenon should be expressed as a hypothesis that generates implications capable of being disconfirmed by empirical tests; and that a hypothesis should be judged successful to the extent that it generates surprising implications that withstand such tests. This approach contrasts with the *model-fitting* methodology used in many previous assessments of level- $k$  theory. The latter methodology works by specifying a parametric level- $k$  model of behaviour in a particular game and by fitting that model to data from (usually experimental) observations. Level- $k$  theory is then judged in terms of the model’s goodness-of-fit, as compared with that of parametric models derived from alternative theories. Many such model-fitting exercises have been carried out and have been interpreted as supporting level- $k$  theory (see Crawford, Costa-Gomes, and Nagore Iriberry, 2012 for a

---

<sup>1</sup> In the closely-related *cognitive hierarchy theory* proposed by Colin Camerer, Teck-Hua Ho, and Juin-Kuan Chong (2004),  $L2$  players choose best replies to a probability mixture of  $L0$  and  $L1$  choices,  $L3$  players choose best replies to a mix of  $L0$ ,  $L1$  and  $L2$ , and so on.

comprehensive review); some of these have used data from games with non-neutral frames (e.g. Crawford and Iriberri, 2007; Crawford, Uri Gneezy and Yuval Rottenstreich, 2008). We will argue that our Popperian approach is better suited to assessing the explanatory power of level- $k$  theory.

We investigate behaviour in a class of two-player simultaneous-move games, defined as follows. For each player  $i = 1, 2$ , there are  $m$  alternative strategies  $s_{i1}, \dots, s_{im}$ , where  $m \geq 3$ .<sup>2</sup> There is a set  $F = \{l_1, \dots, l_m\}$  of distinct *labels*, such that each label  $l_j$  identifies both  $s_{1j}$  and  $s_{2j}$  to both players; this set is the *frame*. Thus, any given strategy  $j$  has the same label for both players. For this reason, we call games in this class *same-label games*. We consider three different payoff matrices for same-label games. A *Hide and Seek* game has the property that if both players choose the same strategy, one player (the *hider*) gets a payoff of 0 and the other (the *seeker*) gets 1; otherwise, the hider gets 1 and the seeker gets 0. In a *Coordination* game, both players get a payoff of 1 if they choose the same strategy and 0 if not. In a *Discoordination* game, they both get 1 if they choose different strategies and 0 if not.

These three games have a common feature that plays an important role in our tests of level- $k$  theory. After abstracting from labelling features, the payoff matrix for each of these games is *strategy-isomorphic* – that is, for each player considered separately, every strategy is isomorphic with every other in that game. This means that within each game, there is no way of distinguishing between the strategies by their payoffs (and hence that uniformly random choice by both players is a Nash equilibrium). Consequently, if players *do* distinguish between strategies in a game, that must be because of differences between labels.<sup>3</sup> If level- $k$  theory is correct, these distinctions must be made by  $L0$  players. (A player at a higher level will take account of labels only if she expects lower-level players to do so.) If  $L0$  players reason non-strategically, as is normally assumed in level- $k$  modelling, there seems no reason to expect their responses to labels to depend on which role (for example, hider or seeker) they are assigned in a strategy-isomorphic same-label game. Equally, there seems no reason to expect  $L0$  responses to labels to differ across games of this kind that have a common frame. By using common frames for the three games, and by assuming that  $L0$

---

<sup>2</sup> Our reason for not including the case where  $m = 2$  is that, following a common practice in the study of Hide and Seek, we want to use frames in which one label is perceived as the unique ‘odd one out’, that is, as being different from all the other labels in some particularly salient respect.

<sup>3</sup> Because strategies are isomorphic, there is no scope for  $L0$  behaviour to be influenced by naïve judgements about the relative payoffs of different strategies, as in the model that Crawford, Gneezy and Rottenstreich (2008) use to explain behaviour in non-neutrally labelled Battle of the Sexes games.

behaviour is the same for coordinators, discoordinators, hiders and seekers, we are able to derive and test level- $k$  hypotheses that apply across the games. This allows much stronger tests than would be possible if each game were investigated separately.

One strength of our tests is that they do not depend on untested assumptions about the relative salience of different labels. In the context of Popperian testing, the hypothesis that  $L0$  players favour ‘salient’ labels is problematic. Although game theorists often appeal to intuitions about salience, there is no generally accepted operational definition of the concept. It is not even clear that the concept has the same meaning in the various contexts in which it is used. Sometimes ‘salience’ is used literally, to refer to a label that a normal person would perceive as standing out from the others in a frame; sometimes it is used to refer to whatever properties of a label induce players to choose it in a Coordination game. A label can be salient in one sense and not in the other (Nicholas Bardsley et al., 2010). However, our design allows us to verify the assumptions that we need to make about  $L0$  players’ choices over labels without addressing questions about whether those labels are ‘really’ salient.

A second strength of our tests is that they do not depend on assumptions about the distribution of player types in the population (except for the general assumption, widely used in level- $k$  modelling, that the relative frequency of  $L0$  is zero). According to level- $k$  theory, the label that is the modal choice at  $L0$  is chosen at different levels by discoordinators, hiders and seekers. This leads to testable implications, independent of the distribution of types, about the *average* frequency with which that label is chosen by players in the three roles. We also derive similarly independent predictions about how, in each of the three games, the choice frequency of that label is affected by changes in the number of labels.

These predictions are derived in Section I. In Section II we explain the experimental design we used to test them. It is based on that used in a series of experiments conducted by Ariel Rubinstein and Amos Tversky (1993), Rubinstein, Tversky and Dana Heller (1996), and Rubinstein (1999), to which we will refer collectively as the work of ‘RTH’. In Section III we report our results. These provide very little support for level- $k$  theory.

In Section IV, we consider how our findings can be reconciled with the apparently conflicting conclusions drawn under the model-fitting methodology. We focus on Crawford and Iriberri (2007; henceforth ‘CI’) because they fit a level- $k$  model to data from some of RTH’s Hide and Seek games. CI treat the goodness-of-fit of this model as evidence of the explanatory power of level- $k$  theory. We show that, viewed in the perspective of classical

statistics, CI’s implicit criterion of explanatory success – the existence of a plausible level- $k$  model that replicates the main patterns in the data – has, by conventional standards, a high probability of false positives. In the final Section, we discuss the wider implications of our results.

## I. THEORY

In this Section, we derive implications of level- $k$  theory for Coordination, Discoordination, and Hide and Seek games that are *matched* with one another – that is, share a common frame.

As explained in the Introduction, we assume that  $L0$  behaviour responds only to labels and hence, for any given frame, is the same for all four player *roles* – coordinator, discoordinator, hider and seeker.<sup>4</sup> We recognise that there *could* be variants of level- $k$  theory in which  $L0$  players respond in some naïve way to the strategic structure of games, and that in such theories,  $L0$  behaviour might differ across roles. But if a theory of this kind is to have predictive power, it must be based on *general* hypotheses about naïve strategic reasoning by  $L0$  players. (If a new  $L0$  assumption were made for every game, the ‘theory’ would be little more than post-hoc rationalisation.) Since such general hypotheses have not been proposed by level- $k$  theorists, our tests assume  $L0$  to be non-strategic.<sup>5</sup>

As described by level- $k$  theory, strategic reasoning takes a similar form for players in all four roles. (At each level above  $L0$ , coordinators and seekers choose the label that is the most frequent choice of co-players at the level below. Discoordinators and hiders choose the label that is the least frequent choice of co-players at the level below.) So there seems to be no reason to suppose that play is more or less sophisticated in any one of these roles than in any other. We therefore assume that (for a given subject pool) the population distribution of types is the same for all four roles. Following CI, we assume that the  $L0$  type has a relative

---

<sup>4</sup> This is a natural generalisation of CI’s assumption that  $L0$  behaviour is the same for hiders and seekers. Although CI do not treat this assumption as essential for a level- $k$  analysis, they describe its neutrality and parsimony as merits of their preferred model (p. 1748).

<sup>5</sup> In some relatives of level- $k$  theory, such as the Hide and Seek model proposed by Michael Bacharach and Stahl (1997), the lowest level of reasoning *described in the model* is naïvely sensitive to strategic features of the game, but the assumed form of naïveté is a best response to a non-strategic opponent. (Bacharach and Stahl assume that hiders avoid salience and seekers favour it. These are best replies to an opponent who non-strategically favours salience.) Apart from the numbering of levels, this is equivalent to assuming a non-strategic  $L0$  type which exists only in the minds of higher-level types.

frequency of zero, and that there are no players at levels higher than  $L4$ . The relative frequencies of the  $L1$ ,  $L2$ ,  $L3$  and  $L4$  types are denoted  $s$ ,  $t$ ,  $u$  and  $v$  respectively.<sup>6</sup>

Our analysis applies to any matched Coordination, Discoordination and Hide and Seek games in which there are  $m$  strategies (with  $m \geq 3$ ) and in which the common frame  $F = \{l_1, \dots, l_m\}$  has an *odd-one-out* property that we now explain. The intuitive idea is that one particular label, denoted (without loss of generality) by  $l_1$ , is perceived by players as standing out from the others, and that those other labels are perceived as equally undistinguished. Thus, in at least one sense of ‘salience’,  $l_1$  is uniquely most salient and  $l_2, \dots, l_m$  are jointly least salient. If the hypothesis that  $L0$  players favour salient labels is interpreted in the same sense, it implies that such players choose  $l_1$  (the *oddity*) with some probability  $q$ , where  $q > 1/m$ , and choose each other  $l_j$  with probability  $(1 - q)/(m - 1)$ . Such a distribution of  $L0$  choices will be called an *odd-one-out distribution* with respect to  $l_1$ .

Initially, we treat the concept of an odd-one-out frame as unambiguous, and derive level- $k$  predictions on the assumption that  $L0$  choices have an odd-one-out distribution with respect to  $l_1$ . But later we will show that if the behaviour of coordinators and discoordination exhibits certain properties, a level- $k$  explanation of that behaviour *must* assume that  $L0$  choices have such a distribution. Thus, for the purposes of a test of level- $k$  theory, observation of those properties can be treated as confirmation of the validity of that assumption, independently of intuitions about salience.

On the initial assumption that players make no errors, and assuming (as CI do) that ties are broken uniformly randomly, the probability with which  $l_1$  is predicted to be chosen by each type in each game is as shown in Table I. (For example, consider the Discoordination game. By assumption,  $L0$  types choose  $l_1$  with probability  $q > 1/m$ .  $L1$  types best-respond to  $L0$  by avoiding  $l_1$ ; since they are indifferent between the other labels, each is chosen with probability  $1/(m - 1)$ .  $L2$  types best-respond to  $L1$  by choosing  $l_1$ . And so on.)

[Table I near here]

---

<sup>6</sup> It is not strictly necessary to assume that there are no players at  $L5$  or above. In all the games we analyse,  $L5$  types would behave just like  $L1$  types,  $L6$  just like  $L2$ , and so on. Thus  $s$  can be reinterpreted as the combined relative frequency of types  $L1$ ,  $L5$ ,  $L9$ , ...;  $t$ ,  $u$  and  $v$  can be reinterpreted similarly. The assumption that the  $L0$  type has zero probability increases the sharpness of the implications we derive, but the main qualitative features of those implications are not critically dependent on it. Allowing a small fraction of  $L0$  types would effectively introduce a noise component that was biased towards the choice of the odd-one-out label.

These theoretical results have the following implications for any set of games with a common odd-one-out frame:

*Implication 1: Coordinators choose the oddity ( $l_1$ ) with probability 1.*

*Implication 2: Discoordinators choose each of the non-oddy labels  $l_2, \dots, l_m$  with equal probability.* This follows from the assumption that  $L0$  players choose each such label with the same probability, and from the assumption of random tie-breaking. The results in Table I then imply that discoordinators choose each non-oddy label with probability  $(s + u)/(m - 1)$ .<sup>7</sup>

*Implication 3: Averaging over equal numbers of discoordinators, hiders and seekers, the oddity is chosen with a probability of at least  $1/3$ .* This follows from the fact that the average of the entries in the ‘all players’ column of Table I for discoordinators, hiders and seekers is  $(s + t + u)/3 + v$  or, equivalently,  $1/3 + 2v/3$  (where  $v \geq 0$ ).

*Implication 4: For each player role (i.e. coordinator, discoordinator, hider, seeker), the probability with which the oddity is chosen is independent of the number of strategies and the frame.* This follows from the fact that, for each role, the ‘all players’ probability depends only on the distribution of types in the population.

*Implication 5: In a population of individuals who repeatedly play the same one of the roles discoordinator, hider, or seeker, each individual either always chooses the oddity or never chooses it.* This follows from the fact that, for each of the three roles, the probability with which the oddity is chosen is 1 for two of the types  $L1, \dots, L4$  and 0 for the other two types; these probabilities are independent of the number of strategies and the frame.

An error structure can be added to the level- $k$  model by relaxing the assumption that players of types  $L1$  and above always know that  $l_1$  is the modal choice of  $L0$  types. Suppose we generalise the model by assuming that, for each higher-level type, the probability that he forms the correct belief about the modal  $L0$  choice is  $1 - e_F$ , where  $0 \leq e_F < [m - 1]/m$ ;<sup>8</sup> if he forms an incorrect belief, each of labels  $l_2, \dots, l_m$  is equally likely to be believed to be modal. Then in every case in which a type tries to replicate the modal  $L0$  choice – that is, every case

---

<sup>7</sup> There are similar implications for hiders (who choose each non-oddy label with probability  $(s + t)/(m - 1)$ ) and for seekers (for whom the probability is  $(t + u)/(m - 1)$ ). We focus on discoordinators because, as we explain later, the predicted effects of deviations from the odd-one-out distribution of  $L0$  choices are less easily detected in the behaviour of hiders and seekers.

<sup>8</sup> This condition is necessary to ensure that  $l_1$  is more likely than any other label to be believed to be modal.



for which the entry in Table I is ‘1’ – the probability that  $l_1$  is chosen is  $1 - e_F$ . For every case in which a type tries to avoid the modal  $LO$  choice – that is, every case for which the entry in Table I is ‘0’ – the probability that  $l_1$  is chosen is  $e_F/(m - 1)$ .<sup>9</sup> Table I then represents the special case in which  $e_F = 0$ . Implication 1 generalises to the proposition that, irrespective of the population distribution of types, the probability with which coordinators choose the oddity is  $1 - e_F$ . Thus, for any given frame  $F$ , players’ propensity to error is revealed in the frequency of non-oddity choices in the relevant Coordination game. If error propensities are known, Implications 2 to 5 can be revised to take account of error. We will explain these revisions in Section III, when we present our results.

Implication 1 can be interpreted as the prediction that oddities act as focal points in coordination games. This is not specific to level- $k$  theory; it is also implied, for example, by theories of team reasoning (Sugden, 1993; Bacharach, 2006; Bardsley et al., 2010). Implication 2 is effectively a re-statement of the assumption that the non-oddity labels are perceived as equally undistinguished; it too is not specific to level- $k$  theory. These implications will be used primarily in verifying the assumption that  $LO$  choices have an odd-one-out distribution. In contrast, Implications 3, 4 and 5 are strong and surprising predictions which, as far as we know, are specific to level- $k$  theory. They are therefore very suitable for Popperian tests of the theory.

## II. EXPERIMENTAL DESIGN

Following the principles explained in Section I, our experiment used matched Coordination, Discoordination and Hide and Seek games to test predictions of level- $k$  theory.<sup>10</sup>

The experiment had two treatments. The *HS treatment* involved 200 subjects, randomly matched into pairs; these pairings were maintained throughout. In the first part of this treatment, each pair played a series of eighteen Hide and Seek games, with no feedback until the end of the experiment. One member of the pair played all eighteen games in the role of hider, the other in the role of seeker; these roles were randomly assigned at the start of the experiment. In the second part of the treatment, the players’ roles were reversed; the same

---

<sup>9</sup> This error structure is formally equivalent to that assumed by CI, except that our error parameter is allowed to depend on the frame. CI’s error parameter  $\varepsilon$  is equivalent to  $e_F m/(m - 1)$  in our notation.

<sup>10</sup> In this respect, our design is similar to that of RTH. Although CI’s level- $k$  model is fitted only to data from Hide and Seek games, RTH’s experiments included some Coordination, Discoordination and Hide and Seek games with common frames.

eighteen games were played, in the same order as before, and again without feedback. After both parts had been completed, one game was selected at random from each part. For each of these games, the winner of that game (i.e. the seeker if both players chose the same item, the hider otherwise) was paid £10, and the loser was paid nothing.

The *CD treatment* involved 80 subjects, randomly matched into pairs, which were maintained throughout. Each pair played a series of eighteen Coordination games, followed by or preceded by a series of eighteen Discoordination games. Which game was played in the first part of the treatment was counterbalanced. There was no feedback until both parts had been completed. Then, for each pair, one game was selected at random from each part. For each of these games, both players were paid £5 if they had achieved the objective of the game (i.e. coordination or discoordination, depending on the game), and nothing otherwise.

Each game was presented to players as a row of either four or eight *boxes*. Each box enclosed a word, symbol or picture. Each player was told that her co-player was seeing the same boxes in the same positions from left to right, and was asked to choose one box, either (in the case of coordinators and seekers) with the aim of choosing the same box as that chosen by her co-player, or (in the case of discoordinators and hidiers) with the aim of choosing a different box. Thus, each box can be interpreted as a distinct strategy label, identified by its content and by its position. Although the positions of the boxes were the same for each member of any given pair of co-players, these positions were independently randomised for each pair and for each game. Thus our design allows the effect of content and position to be investigated independently.<sup>11</sup> In fact, as we report in Section III.1, position had only minor effects on subjects' choices.

Ignoring the randomised positions, each game of a given payoff structure (i.e. Coordination, Discoordination, or Hide and Seek) is defined by a set of four or eight boxes; we will refer to these sets as 'frames'. We used 36 such frames, eighteen with four boxes and eighteen with eight boxes.

Figure I shows the eighteen frames used in eight-box games. These are numbered from 'frame 1b' to 'frame 18b'; each row represents one such frame. For purposes of reference, each box is numbered from left to right in the figure, but of course these numbers have no connection with the positions in which players actually saw the boxes.

---

<sup>11</sup> In RTH's experiments, the 'items' among which subjects chose in any game were arranged in a row, as in our experiment, but their positions were the same for all subjects.

[Figure I near here]

Our intention was that each frame should create an obvious oddity, and that all the other labels should be equally undistinguished. The intended oddity is shown in box 1 in the frames of Figure I. Following Rubinstein et al. (1996), we used two different forms of oddity. In each of frames 1b, 4b, 5b, 6b, 8b and 15b, seven of the boxes are identical; we will say that the box with the distinct content is the *objective oddity*. In each of the other frames, every box is distinct, but one of the boxes is intended to be easily perceived as the *subjective oddity*. Again following Rubinstein et al., the oddity can have *positive*, *negative*, or *neutral* connotations relative to the other boxes in its frame. (The oddity is positive in frames 7b, 8b, 9b, 16b, 17b and 18b, negative in frames 1b, 2b, 3b, 10b, 11b and 12b, and neutral in frames 4b, 5b, 6b, 13b, 14b and 15b.) The main function of this variety of forms of oddity was to maintain subjects' interest and attention. For the purposes of our tests, it is irrelevant what form oddity takes.

The sets of boxes used in four-box games were formed by removing boxes 5, 6, 7 and 8 from each of the sets 1b, ..., 18b, to give the sets 1a, ..., 18a. This method of constructing frames was used to ensure as much similarity as possible between games with different values of  $m$ . Frames 1a, 2a, 4a, 5a, 7a and 9a are virtually identical with those used in the six games investigated by Rubinstein et al.

Table II shows how games were assigned to subjects. In the HS treatment, subjects were randomised into four subgroups (HS1–HS4), each of 50 subjects. In the CD treatment, subjects were randomised into four subgroups (CD1–CD4), each of 20 subjects. Within each subgroup, and within each part of the experiment, the order of the two blocks of nine frames was counterbalanced. The order of games within each block was randomised, independently for each pair. For example, the top part of the first column of the table reports that in the first part of the experiment, each of the 50 subjects in subgroup HS1 played as hider in nine four-box games (frames 1a–9a) and in nine eight-box games (frames 10b–18b). It is evident from the third column that their co-players were the 50 subjects of subgroup HS3.

[Table II near here]

Part 1 of the HS treatment provides 50 observations of hidere and 50 observations of seekers in each of the 36 frames. We will use these data for our main tests of level- $k$  theory, but for completeness we also report the part 2 data. Part 2 of the HS treatment was primarily intended as an add-on investigation of the effects of switching roles. We conjectured that if a

subject played a series of games in one role (hider or seeker) and then played exactly the same games in the other role, there might be some tendency for her choices in the second role to be best responses to her own choices in the first. That is, in terms of level- $k$  theory, behaviour in the second role might be at a higher level than behaviour in the first. For those tests (specifically of Implication 3) that assume the population distribution of levels to be the same for all roles, data from the inexperienced subjects of part 1 of the HS treatment are more suitable than data from part 2.

However, we expected that experience of playing Coordination games would have no significant effect on behaviour in later (and different) Discoordination games. (As we note in Section III.1, this expectation was confirmed.) Thus, the CD treatment was designed with the intention of pooling data from parts 1 and 2, to provide 40 observations of coordinators and 40 observations of discoordinators in each of the 36 frames.

The experiment was conducted in nineteen sessions (eleven for the HS treatment and eight for the CD treatment) at the Centre for Behavioural and Experimental Social Science Laboratory at the University of East Anglia. Subjects were recruited from the general student population and participated anonymously at computer workstations. Instructions were presented on subjects' screens and were also read aloud by an experimenter to ensure that they were common knowledge. In each treatment, subjects were initially given instructions only for the games to be played in the first part of that treatment; instructions for the games to be played in the second part were given only after the first part had been completed. Experimental sessions lasted for approximately 50 minutes. In the CD treatment, average earnings were £7.13; in the HS treatment they were £10 (necessarily, because every Hide and Seek game has a winner and a loser).

### **III. RESULTS**

#### *III.1. Potential confounds: order and position effects*

We begin by eliminating some potentially confounding factors.

The order in which blocks of games were played, and the order in which Coordination and Discoordination games were played in the CD treatment, had no systematic effect on players' behaviour. As we had conjectured might be the case, there were systematic

differences between behaviour in the two parts of the HS treatment.<sup>12</sup> We therefore pool CD data across the two parts of the experiment, but analyse the two parts of the HS data separately.

In all our games, boxes can be distinguished from one another both by content and by position. Because the positions of the boxes were randomised independently for each game and for each pair of co-players, we can test for systematic position effects by comparing the frequencies with which subjects' chosen boxes were in each of the positions 1, ..., 4 (in four-box games) or 1, ..., 8 (in eight-box games). The relevant data are summarised in Table III. It is apparent that, in all cases, each position was chosen with approximately the same frequency. Deviations from equal frequencies are small and show no obvious pattern.

*[Table III near here]*

To test for non-randomness of position choices, we use individual-level data, grouped according to the rows of Table III. For example, take the case of coordinators playing four-box games (the first row of the table), and consider any one position, say position 1. Each of 40 subjects played 18 games in the role of coordinator. For each of these subjects, we find the number of games in which she chose the box in position 1. We then use a chi-squared test to find whether the distribution of these numbers is significantly different from the binomial distribution  $B(18, 0.25)$  implied by random choice. In this particular case, we can reject the null hypothesis at the 5 per cent level. In the 72 tests of this kind (four tests for each of the four-box rows of Table III, eight tests for each of the eight-box rows), we find significant non-randomness ( $p < 0.05$ ) in 17 cases, but with no obvious general pattern. It seems that there may be *some* position effects in our data, but position was not a major determinant of players' choices. From now on, therefore, we abstract from position effects and consider only the content of the boxes in our games and use the terms 'label' and 'frame' to refer only to the content of boxes.

---

<sup>12</sup> Between part 1 and part 2, the frequency of oddity choices, in both four- and eight-box games, increased for hiders and decreased for seekers. Except in the case of eight-box hiders, these differences were statistically significant. Since part 1 seekers tended to favour oddities while part 1 hiders tended to avoid them, these results suggest some tendency for seekers who had previously played as hiders to play best responses to their own previous hiding behaviour, and vice versa. As this issue is orthogonal to our tests of level- $k$  theory, we do not pursue it further.

### III.2. Odd-one-out properties: tests of Implications 1 and 2

With allowance for error, Implication 1 is that the intended oddity (i.e. box 1, defined by content) is the modal choice of coordinators. Table IV reports the frequency with which this item was chosen by coordinators in each of the 18 frames. The contrast with Table III is unmistakable: content was far more important than position in determining players' choices.<sup>13</sup> Notice that, in every case except sets 10b, 11a and 11b, the intended oddity was chosen by more than half the players, and often by very large majorities, even though choices were distributed over four or eight labels. Even in sets 10b and 11b, box 1 was the modal choice. We conclude that, with the exception of Adolf Hitler in set 10b and Kabul in sets 11a and 11b, the intended oddities *were* recognised as uniquely salient by most subjects.<sup>14</sup> Indeed, when we compare the distribution of the numbers of oddity choices made in Coordination games by CD subjects with the binomial distributions implied by random choice,  $B(18, 0.25)$  or  $B(18, 0.125)$  for the four and eight-box games respectively, the difference is overwhelmingly significant ( $p < 0.001$ ). From now on, we will omit the qualifier 'intended', and refer to box 1 as *the* oddity in each set.<sup>15</sup>

[Table IV near here]

Implication 2 is that the non-oddy choices of discoordinators are uniformly distributed. Given that we have abstracted from position effects, this prediction can be tested only for games with subjective oddities. For each of the twelve four-box and twelve eight-box games with subjective oddities, we compare the observed distribution of subjects' choices over non-oddities with the rectangular distribution that is implied by uniformly random choice over non-oddities (given actual choices of oddities). In three of the eight-box games, the expected number of choices of each non-oddy is so small that the expected number of choices of each non-oddy is less than 5, making the chi-squared test unreliable; but none of these games shows any obvious pattern of non-random choice. The null hypothesis is rejected at the 5 per cent level in only one of the 21 reliable tests.

---

<sup>13</sup> The same pattern can be observed in RTH's Coordination games: there was a very strong tendency for subjects to choose the oddity, irrespective of its position.

<sup>14</sup> The first and second most frequently chosen boxes in set 11a were Venice (17 choices), Kabul (10 choices) and Madagascar (also 10 choices). In set 10b they were Hitler (16) and John Lennon (11). In set 11b, they were Kabul (11) and Paris (10).

<sup>15</sup> An alternative method of analysis would be to exclude the data from sets 10b, 11a and 11b. Doing so would make no appreciable difference to our results, except that inferred error rates would be lower.

We conclude that the experimental data are consistent with Implications 1 and 2. If level- $k$  theory is to explain that fact, it must assume that  $L0$  choices have an odd-one-out distribution with respect to  $l_1$ . To see why, first notice that in level- $k$  theory, whichever label is the modal choice at  $L0$  must also be the modal choice of coordinators, and vice versa. So if the data are consistent with Implication 1,  $l_1$  must be the unique modal choice at  $L0$ . Now suppose that  $l_1$  is the modal choice at  $L0$ , and that some other label, say  $l_m$ , is chosen at  $L0$  with a strictly lower probability than every other label. Then, under any plausible assumptions about the population distribution of player types,  $l_m$  will be the unique modal choice of discoordinators, contrary to Implication 2. (At  $L1$ ,  $l_m$  is the unique best response to  $L0$  behaviour.  $L2$  types best-respond to this by randomising over labels other than  $l_m$ . So at  $L3$ ,  $l_m$  is the unique best response to  $L2$  behaviour, and so on.)<sup>16</sup>

### *III.3. Frequency of oddity choices: tests of Implications 3 and 4*

The relative frequencies of oddity choices by players in the four roles are reported in Table V. The second column reports (as percentages) the observed relative frequencies of oddity choices. The first and third columns report the lower and upper bounds of the 95 per cent confidence intervals for these relative frequencies, calculated by treating subjects as the units of observation. (For example, the data in the first row refer to the 40 subjects who played Coordination games. Each subject played 18 such games. The average number of oddity choices by an individual subject over those 18 games is treated as a single observation.)

*[Table V near here]*

Implication 3 is that, in the absence of error, the expected value of the average of the relative frequencies of oddity choices for discoordinators, hiders and seekers is at least 0.333. (We will call this the *synthetic average*.) For the reasons explained in Section II, we use only the part 1 choices of hiders and seekers for this test (although in fact our main conclusions would be unchanged if we used part 2 data instead). The observed synthetic average for four-box games is only 0.212. For eight-box games, the observed synthetic average is still lower, at 0.164. Even if, for each of the three roles, we were to use the upper bound of the relevant confidence interval rather than the observed value, the synthetic average would still be only

---

<sup>16</sup> For hiders and seekers, the predicted effects of a uniquely undistinguished label  $l_m$  are less sharp. Overall, hiders and seekers will favour *both*  $l_1$  and  $l_m$  (because  $l_1$  is chosen by  $L1$  seekers and  $L2$  hiders, while  $l_m$  is chosen by  $L1$  hiders and  $L4$  seekers). For this reason, we use discoordinators' behaviour to test for the absence of uniquely undistinguished labels.

0.252 for four-box games and 0.209 for eight-box games. So the hypothesis that the synthetic average is at least 0.333 can be rejected with confidence.

This conclusion is not affected if errors are taken into account. Implication 3 generalises to the prediction that the expected value of the synthetic average is at least  $(1/3)(1 - e_F) + (2/3)(e_F/[m - 1])$ . Recall that the value of the error parameter  $e_F$  can be inferred from the frequency of oddity choices by coordinators. Averaging over all four-box games, coordinators chose the oddity with an average relative frequency of 0.76. Exactly the same relative frequency was observed in eight-box games.<sup>17</sup> The implied value of  $e_F$  (for both four- and eight-box frames) is therefore 0.24. Using this value, the synthetic average is predicted to be at least 0.306 for four-box games and at least 0.276 for eight-box games – still well outside the confidence intervals of our observations. We conclude that our observations are not consistent with Implication 3. In simple terms, level- $k$  theory predicts that the choices of discoordinators, hiders and seekers will be skewed towards the oddity to a degree that we do not observe.

Implication 4 is that, in the absence of error and for any given player role, the expected proportion of oddity choices is the same for four- and eight-box games. For consistency with our tests of Implication 3, the tests we report use only part 1 data for hiders and seekers. (In fact, the part 2 data give even less support for Implication 4.) The observed proportions of oddity choices at the two values of  $m$  are exactly equal for coordinators and almost exactly equal for seekers, but for discoordinators and hiders, the four-box proportions are much greater than the corresponding eight-box proportions. These differences are significant at the 5 per cent level (as can be seen from the confidence intervals). Of course, some reduction in the frequency of oddity choices in moving from four-box to eight-box games can be attributed to error, but it seems unlikely that this effect is sufficient to account for the differences we observe. Setting  $e_F = 0.24$ , level- $k$  theory implies that the margin by which the four-box proportion of oddity choices exceeds the eight-box proportion is  $0.046(s + u)$  for discoordinators,  $0.046(s + t)$  for hiders, and  $0.046(t + u)$  for seekers.<sup>18</sup> So we might

---

<sup>17</sup> We conjecture that this equality is the result of two opposing effects offsetting one another. On the one hand, the more boxes there are to choose from, the more candidates there are for the status of ‘most salient box’. On the other, the unique property of an oddity is more obvious, the more items lacking that property are presented alongside it.

<sup>18</sup> In general, the predicted margin for discoordinators is  $e_F([1/3] - [1/7])(s + u)$ , which equals  $0.046(s + u)$  when  $e_F = 0.24$ . A similar analysis applies to hiders and seekers.



expect error to induce margins of the order of 0.02 to 0.03. The observed margins are 0.079 for discoordinators, 0.060 for hidere, and 0.005 for seekers.

#### *III.4. Consistency at the individual level: tests of Implication 5*

We now use individual-level data to test Implication 5 – that, in the absence of error, an individual who repeatedly plays as a discoordinator, a hider or a seeker will either always choose the oddity or never choose it. We begin with the case of discoordinators in four-box games.

According to the level- $k$  model, there are two *oddity-choosing* types in the Discoordination game, namely  $L2$  and  $L4$ . A subject who is of one of these types chooses the oddity in every game, except as a result of error. Other subjects choose non-oddities in every game, except as a result of error (see Table I). Allowing for error, oddity-choosing types choose the oddity in each game with an independent probability of  $1 - e_F$ ; other types choose it with an independent probability of  $e_F/3$ . Setting  $e_F = 0.24$ , these probabilities are 0.76 and 0.08 respectively. In the experiment, 40 subjects each faced 18 discoordination games. We notionally divide these subjects between oddity-choosing and non-oddity-choosing types so that the expected proportion of oddity choices is as close as possible to the observed proportion (0.233). We construct an ‘expected’ distribution of ‘number of oddities chosen by subject’ by summing the binomial distributions appropriate to the two categories of subject. We then compare this with the observed distribution. These distributions are plotted in the first panel of Figure II. It is immediately obvious that the two distributions are very different, contrary to Implication 5. The expected distribution is bimodal, reflecting the assumption that subjects belong to two distinct categories with different behaviour patterns. The observed distribution is much closer to the unimodal binomial distribution one would expect if every subject randomised her choices across the experiment, with a common and constant probability of choosing the oddity in each task. The hypothesis that the observations are drawn from the predicted distribution is strongly rejected by a chi-squared test ( $p < 0.01$ ).

We carried out the same test for eight-box discoordinators, and for hidere and seekers (using part 1 data). For hidere, the oddity-choosing types are  $L3$  and  $L4$ ; for seekers, they are  $L1$  and  $L4$ . In every case, as Figure II shows, the observed distribution differs from the predicted distribution in the general direction we have described for four-box discoordinators. In every case, the hypothesis derived from Implication 5 is strongly rejected ( $p < 0.01$ ). These results give little support to the hypothesis, fundamental to level- $k$  theory, that

differences between the strategy choices of different players in a given game can be explained in terms of between-player differences in levels of reasoning. It seems that the main source of these differences is within-player stochastic variation.

#### IV. A COMPARISON WITH THE MODEL-FITTING APPROACH

The tests reported in Section III provide very little support for level- $k$  theory. In this respect, our findings contrast with the conclusions that CI draw from fitting a level- $k$  model to data from RTH's experiments. This may seem surprising, particularly as our experimental design was based on RTH's and used many of the same frames. In this Section, we investigate this contrast.

CI claim that RTH's Hide and Seek data show 'robust patterns', namely that the strategy with the *least* salient label 'was the strongly modal choice for both hiders and seekers, and was even more prevalent for seekers than hiders' (p. 1732). We will call the combination of these two properties the *fatal attraction pattern*. (The tendency of hiders to choose the least salient label, despite seekers having the same tendency, is the 'fatal attraction' in the title of CI's paper.) CI show that a level- $k$  model can be fitted to these data. On the strength of this, they claim that level- $k$  theory can provide a 'simple explanation of RTH's and related results' (p. 1734).

CI's model-fitting exercise can be separated into two stages. The first stage is the construction of a data set that exhibits the fatal attraction pattern. This is not merely a matter of reassembling RTH's experimental data: judgements are made about which data to include and how data should be classified. The main reason why these judgements are required is that the pattern to be explained is defined in terms of 'salience', and that concept is not directly observable, independently of model-fitting criteria. In order to include any given RTH game in their data set, CI have to make preliminary judgements about which of its labels are to be treated as most and least salient. We will argue that, because of the subjectivity of these judgements, the robustness of the fatal attraction pattern is open to question.

The second stage of the model-fitting exercise is to specify a variety of parametric models and to fit these to the constructed data set. CI propose a particular level- $k$  model on grounds of plausibility and goodness of fit; the fact that this model replicates the fatal

attraction pattern is treated as evidence of the explanatory power of level- $k$  theory.<sup>19</sup> We will show that, even given the premise that the pattern exists, the existence of a plausible level- $k$  model that replicates the pattern is only weak evidence of explanatory power.

In constructing their data set, CI begin with a theoretical analysis of one of RTH's Hide and Seek games, which is treated as a baseline case. The frame for this game is the same as our frame 4a; CI call it an 'ABAA' game. Subjects are shown four 'items' arranged in a row from left to right – the letters A, B, A, A. The hider chooses an item behind which to hide a 'treasure'; the seeker guesses which item is the location of the treasure. CI's assumptions about  $L0$  behaviour are derived from the following argument about salience: 'The "B" location is distinguished by its label, and so is salient in one of Thomas Schelling's (1960) senses. And the two "end A" locations, though not distinguished by their labels, may be inherently salient, as [Rubinstein and Tversky (1993) and Rubinstein et al. (1996)] argue, citing Nicholas Christenfeld (1995).' On the basis of this reasoning, and 'because the end A frequencies are almost equal in the data' (p. 1738), CI assume that 'central A' is the least salient label (p. 1732) and that the two end As are equally salient. These assumptions allow two alternative cases. In *Case 1*, B is uniquely most salient. In *Case 2*, the end As are jointly most salient. In other words,  $L0$  players (whether hiders or seekers) choose A, B, A, A (from left to right) with probabilities  $p/2$ ,  $q$ ,  $1 - p - q$  and  $p/2$  respectively, where  $p > 1/2$  and  $q > 1/4$ ; Case 1 applies if  $p < 2q$  and Case 2 applies if that inequality is reversed (p. 1738).

While the unique salience of B in Case 1 seems intuitively uncontroversial (and it corresponds to what we have found), Case 2 is more problematic, since Christenfeld's finding is that when individuals pick from a row of identical items, they tend to *avoid* the end locations. On a natural interpretation, that finding implies that end locations are *not* salient (i.e. are not favoured by non-strategic agents). A further problem, common to both cases, concerns the assumption that the two end As are chosen with equal frequency by  $L0$  types. This forces CI's level- $k$  model to predict that they are chosen with equal frequency by *all* types. Since this assumption has been justified by the observation of equal frequencies in RTH's data, it cannot also be used to *explain* that observation. The implication is that CI's model can be used to explain the distributions of hiders' and seekers' choices between only

---

<sup>19</sup> In a further stage of analysis, CI test the 'portability' of their estimated model by applying it to experimental data from two different games with non-neutral frames. However, the frames of these games are too different from RTH's Hide and Seek games to allow portability of the  $L0$  specification. In effect, CI use a new  $L0$  specification for each of the games; all that is carried over from the RTH analysis is the population distribution of types and the error parameter.

three categories: B, central A, and *the combination of the two end As*. Thus, the data that the model is to explain consist of three relative frequencies of choices for each of two roles. Since for each role the relative frequencies must sum to one, there are only four degrees of freedom.

The data analysed by CI come from six Hide and Seek games selected from RTH's experiments. Two of these are implementations of the baseline game in different experiments. A third differs only in that B appears in the third position rather than the second; reasonably enough, CI treat the A in the second position in the AABA frame as equivalent to the 'central A' of the baseline game. A fourth game has the same frame and the same matrix as the third game, but is described differently to players. One player (corresponding to the seeker of the treasure in the baseline game) chooses a location to place a 'mine', with the objective that the other player will *choose* this location; the other player (corresponding with the hider of the treasure) guesses a location in which the mine has *not* been placed. If *L0* behaviour is assumed to be the same for seekers as for hiders, the equivalence between treasure and mine games can be interpreted as an implication of level-*k* theory.<sup>20</sup>

The remaining two games are the treasure and mine versions of a Hide and Seek game in which each player chooses one of '1', '2', '3' or '4'. Here CI's arguments are less compelling. They treat 1 and 4 as analogous with the 'end As' in the baseline game and 2 as analogous with 'B', thus classifying 3 as 'least salient'. The only explanation they offer is that they are following Rubinstein and Tversky's (1993: 4) suggestion that 'the least salient response ... may correspond to 3, or perhaps 2' (CI, 2007: 1736). With the assumption that it is 3, this creates two additional instances of the fatal attraction pattern.

Three games reported by Rubinstein et al. (1996) might be thought similar to ABAA games, but are not included in CI's analysis: the frames for these games are essentially the same as our frames 2a, 7a, and 8a, but with the odd-one-out in second or third position.<sup>21</sup> CI (private communication) have told us that they excluded these games because the affective quality of the odd-one-out was a confounding factor. That is a defensible judgement call, but

---

<sup>20</sup> However, CI's motivation for this assumption seems to appeal to the fact that in RTH's experiments 'the mine treatments yielded results very close to the treasure treatments with the roles reversed' (p. 1736). Such an appeal is illegitimate if CI's objective is to *explain* RTH's results.

<sup>21</sup> The other two games reported by Rubinstein et al. are less like ABAA games, because the odd-one-out is either first or last in the row.

had these games been included there would have been two exceptions to the fatal attraction pattern: the modal choice for both hiders and seekers was the odd-one-out in the RTH versions of 7a and 8a, as in our experiment.

Having constructed their data set in the way we have described, CI are able to report that behaviour in all six games has the fatal attraction pattern. Finding no significant differences between the distributions of choices across the games, given their classification system, they pool the data and treat them as if they had all been generated by the baseline game. This gives the  $2 \times 3$  array of relative frequencies of choice to which their models are fitted.

CI's preferred class of level- $k$  models assume: (i)  $L0$  behaviour is the same for hiders and seekers; (ii) the relative frequencies of  $L1$ ,  $L2$ ,  $L3$  and  $L4$  types are  $s$ ,  $t$ ,  $u$  and  $v$ , where  $s + t + u + v = 1$ ; and (iii)  $p > 1/2$  and  $q > 1/4$ . Assumptions (i) and (ii) were explained in Section I; the salience assumption (iii) is common to Cases 1 and 2. Within this class, the best-fitting model has the Case 2 form:  $p > 2q$  (i.e. in terms of the baseline game, the end As are more salient than B). The estimated parameter values are  $s = 0.19$ ,  $t = 0.32$ ,  $u = 0.24$ ,  $v = 0.25$  (CI, Table III). CI present this as their 'proposed model' and as a 'convincing' explanation of RTH's results (p. 1748).

In evaluating this claim, it is useful to ask the following question. Let us accept that RTH's data have the fatal attraction pattern. A plausible level- $k$  model generates that pattern too. Is that coincidence sufficiently surprising that it can be treated as evidence of the explanatory power of level- $k$  theory? More precisely: How likely is it that, *given any arbitrary pattern in the data*, that pattern would be implied by *some* plausible level- $k$  model?

To investigate this issue, we need a general definition of a 'pattern', such that the fatal attraction pattern is an instance; and we need a definition of 'plausibility'. Consider a large population of potential players of the baseline Hide and Seek game. Let  $B_H$  and  $B_S$  be the probabilities with which hiders and seekers choose B. Let  $C_H$  and  $C_S$  be the probabilities with which hiders and seekers choose 'central A'. We define  $E_H = (1 - B_H - C_H)/2$  and  $E_S = (1 - B_S - C_S)/2$ . That is,  $E_H$  and  $E_S$  are the probabilities with which hiders and seekers choose each of the 'end A' labels. Assuming that there are no ties, we define  $m_H, m_S \in \{B, C, E\}$  as the *modal choices* of hiders and seekers. (Thus, for example,  $m_H = B$  denotes ' $B_H > C_H$  and  $B_H > E_H$ '.) We define a variable  $z \in \{H, S\}$  such that  $z = H$  if the modal choice of hiders is chosen with greater probability than the modal choice of seekers;  $z = S$  denotes the converse.

Finally, we define a *pattern* as a triple  $(m_H, m_S, z)$ . There are eighteen possible patterns. One of these, namely  $(C, C, S)$  is the fatal attraction pattern.

Following what we take to be CI's intentions, we define a level- $k$  model to be *plausible* if it satisfies their assumptions (i) to (iii) and also has the properties  $t > s$  and  $u > v$  (i.e. the distribution of levels is 'hump-shaped'). CI say that this is the shape of distribution most commonly estimated in level- $k$  models (p. 1734); they describe their preferred model as 'behaviorally plausible' because, among other features, it has 'the characteristic hump-shape of previous estimates' (p. 1743).

We can now ask how many of the eighteen possible patterns are implied by plausible level- $k$  models. The answer, which we derive in the Appendix, is seven – that is, 39 per cent of the total. In other words, if the experimental data were generated entirely at random, the probability that they would be capable of being 'explained' by a plausible level- $k$  model is 0.39. This is far higher than the conventional significance levels used in hypothesis-testing. So the fact that a plausible level- $k$  model can 'explain' the pattern in RTH's data is only weak evidence in support of the theory.

To summarise, we suggest that the key source of the apparent contrast between our conclusions and CI's is in our respective responses to the absence of a received theory of the determinants of salience. A level- $k$  model of Hide and Seek needs a specification of how  $L0$  players distribute their choices among strategies that are isomorphic with respect to payoffs but differ in terms of labels. It might seem that this requires the modeller *either* to make specific assumptions about salience, justified by appeals to intuition or to contestable theoretical hypotheses, *or* to allow the data to decide which of a set of alternative  $L0$  specifications gives the best fit. The first approach makes it difficult to distinguish between successful model-fitting and post-hoc rationalisation. The second approach increases the number of free parameters, reducing the parsimony of the model. CI steer a middle course, imposing some specific assumptions about salience but fitting two alternative  $L0$  specifications and choosing between those in terms of goodness of fit. As we have shown, some of those assumptions are contestable; and even if they are accepted, the resulting test is statistically too weak to allow any strong conclusion to be drawn.

Our experimental design reflects a very different response to the absence of a received theory of salience. By investigating Coordination, Dicoordination and Hide and Seek games with common frames, we can employ statistically powerful tests that do not

depend on any prior assumptions about salience. If those tests reject hypotheses that are consistent with model-fitting exercises, that is not so surprising.

## V. CONCLUSION

Following a Popperian strategy, we have derived predictions from level- $k$  theory and have subjected these to experimental tests. In particular, we have focussed on three intuitively surprising predictions about behaviour in Coordination, Discoordination and Hide and Seek games with common odd-one-out frames. First, the theory predicts a general tendency for the choices of discoordinators, hidiers and seekers to favour oddities. Second, this tendency is predicted to be stronger, the larger the number of strategies in the relevant game. Third, when individuals play more than one game as discoordinator, as hider, or as seeker, each individual is predicted *either* consistently to choose the oddity with probability close to 1 *or* consistently to choose it with probability close to 0. Our experiment had the capacity to detect these three effects, were they to occur. We found very little evidence that they did.

We cannot think of any plausible variant of level- $k$  theory that would not have these predictions. The implications of any level- $k$  model are fully determined by the specification of  $L0$  behaviour, by the population distribution of types, and by the error structure. The assumptions that we have made about  $L0$  behaviour in games with odd-one-out frames are verified by our observations of behaviour in Coordination and Discoordination games. That  $L0$  behaviour is the same in all four roles is an implication of the idea, fundamental to level- $k$  theory's claim to explanatory power, that  $L0$  reasoning is non-strategic. Given that these assumptions are satisfied, the three predictions are independent of how 'salience' is defined and explained, independent of the population distribution of types, and robust to plausible degrees of error.

Some readers may expect us to end by offering conjectures about how our results might be explained, if the explanation offered by level- $k$  theory were rejected. However, this is not in the nature of a Popperian approach. A Popperian experimenter tries to test predictions that are unique to the theory under consideration. The more surprising those predictions are, when viewed in the perspectives of competing theories, the more powerful such a test is. But that is to say that, viewed in the perspectives of those other theories, disconfirmations of the predictions are *unsurprising* and therefore *uninformative*. For this reason, a test of one theory should not be expected to provide information that is particularly useful for assessing competing theories. In the present case, the evidence from our

experiment and from those of Rubinstein, Tversky and Heller shows that, for some frames and for some subject pools, behaviour in Hide and Seek games is influenced by labels.<sup>22</sup> Characterising and explaining such effects is a significant problem for behavioural game theory. But our experiment was not designed to collect the kind of data required for a systematic attack on that problem. It was designed to test, and in the event rejected, *one particular hypothesis* about the cause of those effects.

We conclude instead with an observation about focal points. In disconfirming predictions of level- $k$  theory that relate behaviour in Coordination games to behaviour in other games with non-neutral frames, we have posed problems for that theory's explanation of how players use labels to identify focal points. According to that explanation, the distinguishing feature of a focal point is that its label tends to be chosen by the most strategically naïve players. Interestingly, Schelling (1960: 94, italics in original) rejects that explanation in his discussion of John Maynard Keynes's (1936: 156) 'beauty contest' game, in which each player's aim is to choose, from a set of a hundred photographs, the six faces that the average player deems to be 'prettiest'. Keynes's analysis is now widely recognised as an anticipation of level- $k$  theory. In modern language, the essential idea is that the least sophisticated players are non-strategic  $L0$  types who choose the faces that they personally find prettiest; more sophisticated players reason at higher levels. Schelling recognises that Keynes's analysis could be adapted to generate an explanation of focal points, but says that this 'is *not* at all the same' as his own theory, in which players of a Coordination game jointly look for a label that is prominent when considered as a potential solution to their coordination problem.<sup>23</sup> Our results provide some support for Schelling's scepticism about the level- $k$  approach.

SCHOOL OF ECONOMICS, UNIVERSITY OF EAST ANGLIA (HARGREAVES HEAP AND SUGDEN);  
ECONOMIC SCIENCE INSTITUTE, CHAPMAN UNIVERSITY (ROJO ARJONA)

---

<sup>22</sup> Whether this is true for Discoordination games is less clear. In our experiment, discoordinators' behaviour was consistent with the hypothesis of uniformly random choice. However, Rubinstein et al. (1993) report two (out of six) Discoordination games in which choices were significantly non-random; in one of these games the modal choice was the odd-one-out, but in the other it was not.

<sup>23</sup> The theory of team reasoning can be interpreted as an attempt to formalise Schelling's analysis. Experimental tests of the team-reasoning explanation of focal points, like those of the level- $k$  explanation, have had mixed results (e.g. Crawford, Uri Gneezy, and Yuval Rottenstreich, 2008; Bardsley et al., 2010).



## APPENDIX: PATTERNS CONSISTENT WITH ‘PLAUSIBLE’ LEVEL-K MODELS

This Appendix uses the same notation and assumptions as are used in Sections I and IV of the main paper. We treat the following assumptions about the population distribution of types as necessary properties of a ‘plausible’ model:

- (1)  $s, t, u, v \geq 0$
- (2)  $s + t + u + v = 1$
- (3)  $t > s$
- (4)  $u > v$ .

Two alternative assumptions about salience are treated as ‘plausible’: *either* ‘B’ is most salient (Case 1), *or* the endpoints are most salient (Case 2). We consider these in turn.

### Case 1: $p < 2q$

Level- $k$  theory implies the following choice probabilities:

- (5)  $B_H = u/3 + v$
- (6)  $C_H = s + t/3$
- (7)  $E_H = t/3 + u/3$
- (8)  $B_S = s + v/3$
- (9)  $C_S = t + u/3$
- (10)  $E_S = u/3 + v/3$

Expressions (1) to (10) imply the following results:

*Result 1(i):* Not  $[m_S = B]$ . *Proof:* Suppose  $m_S = B$ . Then  $B_S > C_S$  which, by (8) and (9), implies  $s + v/3 > t + u/3$ . But by (3) and (4),  $s + v/3 < t + u/3$ , a contradiction.

*Result 1(ii):* Not  $[m_H = C \text{ and } m_S = E]$ . *Proof:* Suppose  $m_H = C$  and  $m_S = E$ . Then  $C_H > E_H$  which, by (6) and (7) implies  $u/3 < s$ . But also  $E_S > B_S$  which, by (8) and (10), implies  $s < u/3$ , a contradiction.

*Result 1(iii):* Not  $[m_H = E \text{ and } m_S = E]$ . *Proof:* Suppose  $m_H = E$  and  $m_S = E$ . Then  $E_H > B_H$  which, by (5) and (7), implies  $t/3 > v$ . But also  $E_S > C_S$  which, by (9) and (10), implies  $v/3 > t$ . Adding these inequalities gives  $(t + v)/3 > t + v$ . Given (1), this is a contradiction.

*Result 1(iv):* Not  $[m_S = E \text{ and } z = S]$ . *Proof:* suppose  $[m_S = E \text{ and } z = S]$ . Then  $E_S > B_H, C_H, E_H$ . But by (5) and (11),  $B_H - E_S = 2v/3$ , which by (1) is non-negative: a contradiction.

*Result 1(v):* Not  $[m_H = E \text{ and } z = H]$ . *Proof:* suppose  $[m_H = E \text{ and } z = H]$ . Then  $E_H > B_S, C_S, E_S$ . But by (7) and (9),  $C_S - E_H = 2t/3$ , which by (1) is non-negative: a contradiction.

Of the eighteen possible patterns, only six are consistent with Results 1(i) to 1(v). These are listed below. For each pattern, we give an example of a ‘plausible’ distribution of types that induces that pattern.

Pattern 1:  $(B, C, H)$ . Example:  $s = 0.05, t = 0.15, u = 0.51, v = 0.29$ .

Pattern 2:  $(B, C, S)$ . Example:  $s = 0.05, t = 0.36, u = 0.30, v = 0.29$ .

Pattern 3:  $(B, E, H)$ . Example:  $s = 0.04, t = 0.06, u = 0.50, v = 0.40$ .

Pattern 4:  $(C, C, H)$ . Example:  $s = 0.35, t = 0.40, u = 0.15, v = 0.10$ .

Pattern 5:  $(C, C, S)$ . Example:  $s = 0.20, t = 0.40, u = 0.30, v = 0.10$ .

Pattern 6:  $(E, C, S)$ . Example:  $s = 0.10, t = 0.13, u = 0.73, v = 0.04$ .

## Case 2: $p > 2q$

Level- $k$  theory implies the following choice probabilities:

$$(11) \quad B_H = t/2 + u/3$$

$$(12) \quad C_H = s + t/2$$

$$(13) \quad E_H = u/3 + v/2$$

$$(14) \quad B_S = u/2 + v/3$$

$$(15) \quad C_S = t + u/2$$

$$(16) \quad E_S = s/2 + v/3$$

Expressions (1) to (4) and (11) to (16) imply the following results:

*Result 2(i):* Not  $m_S = E$ . *Proof:* suppose  $m_S = E$ . Then  $E_S > C_S$  which, by (15) and (16), implies  $s/2 + v/3 > t + u/2$ . But by (1), (3) and (4),  $0 \leq s < t$  and  $0 \leq v < u$ . Thus  $s/2 < t$  and  $v/3 < u/2$ . Summing these inequalities gives  $s/2 + v/3 < t + u/2$ , a contradiction.

*Result 2(ii):* Not  $[m_H = B \text{ and } m_S = B]$ . *Proof:* Suppose  $m_H = B$  and  $m_S = B$ . Then  $B_H > E_H$  which, by (11) and (13), implies  $t > v$ . But also  $B_S > C_S$  which, by (14) and (15), implies  $v/3 > t$ . Given (1), this implies  $v > t$ , a contradiction.

*Result 2(iii):* Not  $[m_H = C \text{ and } m_S = B]$ . *Proof:* Suppose  $m_H = C$  and  $m_S = B$ . Then  $C_H > B_H$  which, by (11) and (12), implies  $s > u/3$ . But also  $B_S > C_S$  which, by (14) and (15), implies  $v/3 > t$ . Adding these inequalities gives  $s + v/3 > u/3 + t$ . But by (3) and (4),  $s < t$  and  $v < u$ , a contradiction.

*Result 2(iv):* Not  $[m_H = E \text{ and } z = H]$ . *Proof:* suppose  $[m_H = E \text{ and } z = H]$ . Then  $E_H > B_S, C_S, E_S$ . But by (13) and (14),  $B_S - E_H = (u - v)/6$ , which by (4) is positive: a contradiction.

Of the eighteen possible patterns, only six are consistent with Results 2(i) to 2(iv). These are listed below. For the one pattern that is not also consistent with Case 1, we given an example of a ‘plausible’ distribution of types that induces that pattern:

Pattern 1:  $(B, C, H)$ . Also consistent with Case 1.

Pattern 2:  $(B, C, S)$ . Also consistent with Case 1.

Pattern 4:  $(C, C, H)$ . Also consistent with Case 1.

Pattern 5:  $(C, C, S)$ . Also consistent with Case 1.

Pattern 6:  $(E, C, S)$ . Also consistent with Case 1.

Pattern 7:  $(E, B, S)$ . Example:  $s = 0.04, t = 0.06, u = 0.50, v = 0.40$ .

We have now established that exactly seven of the eighteen possible patterns are consistent with ‘plausible’ level- $k$  models.

## REFERENCES

- Bacharach, Michael. 2006. *Beyond Individual Choice: Teams and Frames in Game Theory*. Princeton, NJ: Princeton University Press.
- Bacharach, Michael and Dale O. Stahl. 1997. "Variable-Frame Level- $n$  Theory." <http://www.eco.utexas.edu/~stahl/vflnt4a.pdf>.
- Bardsley, Nicholas, Judith Mehta, Chris Starmer and Robert Sugden. 2010. "Explaining Focal Points: Cognitive Hierarchy Theory *versus* Team Reasoning." *Economic Journal* 120: 40–79.
- Camerer, Colin F., Teck-Hua Ho, and Juin-Kuan Chong. 2004. "A Cognitive Hierarchy Model of Games." *Quarterly Journal of Economics* 119(3): 861–898.
- Christenfeld, Nicholas. 1995. "Choices from Identical Options." *Psychological Science* 6(1): 50–55.
- Costa-Gomes, Miguel A., Vincent P. Crawford and Bruno Broseta. 2001. "Cognition and Behavior in Normal-Form Games: An Experimental Study." *Econometrica* 69(5): 1193–1235.
- Crawford, Vincent P., Miguel A. Costa-Gomes, and Nagore Iriberri. 2012. "Structural Models of Nonequilibrium Strategic Thinking: Theory, Evidence, and Applications." Forthcoming in *Journal of Economic Literature*.
- Crawford, Vincent P., Uri Gneezy, and Yuval Rottenstreich. 2008. "The Power of Focal Points Is Limited: Even Minute Payoff Asymmetry May Yield Large Coordination Failures." *American Economic Review*, 98 (4): 1443–1458.
- Crawford, Vincent P. and Nagore Iriberri. 2007. "Fatal Attraction: Salience, Naïveté, and Sophistication in Experimental Hide-and-Seek Games." *American Economic Review* 97(5): 1731–1750.
- Keynes, John Maynard. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.
- Nagel, Rosemarie. 1995. "Unraveling in Guessing Games: An Experimental Study." *American Economic Review* 85(5): 1313–1326.
- Rubinstein, Ariel. 1999. "Experience from a Course in Game Theory: Pre and Post-Class Problem Sets as a Didactic Device." *Games and Economic Behavior* 28(1): 155–170.

Rubinstein, Ariel and Amos Tversky. 1993. "Naïve Strategies in Zero-Sum Games." Working Paper 17- 93, Sackler Institute of Economic Studies, Tel Aviv University.

Rubinstein, Ariel, Amos Tversky, and Dana Heller. 1996. "Naïve Strategies in Competitive Games." In *Understanding Strategic Interaction: Essays in Honor of Reinhard Selten*, ed. Wulf Albers, Werner Güth, Peter Hammerstein, Bemmy Moldovanu, and Eric van Damme, 394–402. Berlin: Springer-Verlag.

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

Stahl, Dale O. and Paul Wilson. 1994. "Experimental Evidence On Players' Models of Other Players." *Journal of Economic Behavior and Organization* 25(3): 309–327.

Sugden, Robert. 1993. "Thinking as a Team: Towards an Explanation of Non-Selfish Behavior." *Social Philosophy and Policy* 10: 69-89.

**TABLE I: IMPLICATIONS OF LEVEL- $k$  THEORY FOR MATCHED ODD-ONE-OUT GAMES  
(WITHOUT ERRORS)**

game	choice probability for $l_1$ :					all players
	$L0$	$L1$	$L2$	$L3$	$L4$	
coordinators	$q$	1	1	1	1	1
discoordinators	$q$	0	1	0	1	$t + v$
hiders	$q$	0	0	1	1	$u + v$
seekers	$q$	1	0	0	1	$s + v$

**TABLE II: ASSIGNMENT OF GAMES TO SUBJECTS**

	subgroup (and number of subjects)							
	HS1	HS2	HS3	HS4	CD1	CD2	CD3	CD4
	(50)	(50)	(50)	(50)	(20)	(20)	(20)	(20)
<i>Part 1</i>								
frames 1a–9a	H	–	S	–	C	–	D	–
frames 1b–9b	–	H	–	S	–	C	–	D
frames 10a–18a	–	H	–	S	C	–	D	–
frames 10b–18b	H	–	S	–	–	C	–	D
<i>Part 2</i>								
frames 1a–9a	S	–	H	–	–	D	–	C
frames 1b–9b	–	S	–	H	D	–	C	–
frames 10a–18a	–	S	–	H	–	D	–	C
frames 10b–18b	S	–	H	–	D	–	C	–

Letters denote roles played (H = hider, S = seeker, C = coordinator, D = discoordinator).

**TABLE III: FREQUENCY OF CHOICES BY POSITION**

role	percentage of choices that are of position:							
	1	2	3	4	5	6	7	8
<i>four- box games</i>								
coordinators ( $n = 40$ )	27.5	24.4	26.1	21.9				
discoordinators ( $n = 40$ )	26.4	20.3	26.8	26.5				
hiders (part 1: $n = 100$ )	25.0	22.2	24.9	27.9				
hiders (part 2: $n = 100$ )	24.3	24.6	24.9	26.2				
seekers (part 1: $n = 100$ )	22.9	27.7	28.3	21.1				
seekers (part 2: $n = 100$ )	22.2	25.9	31.2	20.7				
<i>eight- box games</i>								
coordinators ( $n = 40$ )	10.2	11.8	8.0	9.7	13.6	9.2	9.8	7.8
discoordinators ( $n = 40$ )	8.1	10.8	12.8	8.0	10.1	10.1	11.9	8.2
hiders (part 1: $n = 100$ )	12.1	13.0	13.0	10.2	9.8	13.3	15.6	13.0
hiders (part 2: $n = 100$ )	13.7	14.2	11.7	11.1	10.8	14.4	12.9	11.2
seekers (part 1: $n = 100$ )	10.0	13.4	13.8	13.2	13.6	12.8	12.3	10.9
seekers (part 2: $n = 100$ )	10.2	10.7	15.8	14.0	15.4	14.4	11.2	8.2

Note:  $n$  denotes the number of subjects who faced games of the relevant type. Each coordinator and discoordinator faced 18 games of that type; each hider and seeker faced 9 games of that type.



**TABLE IV: FREQUENCY OF ODDITY CHOICES IN COORDINATION GAMES**

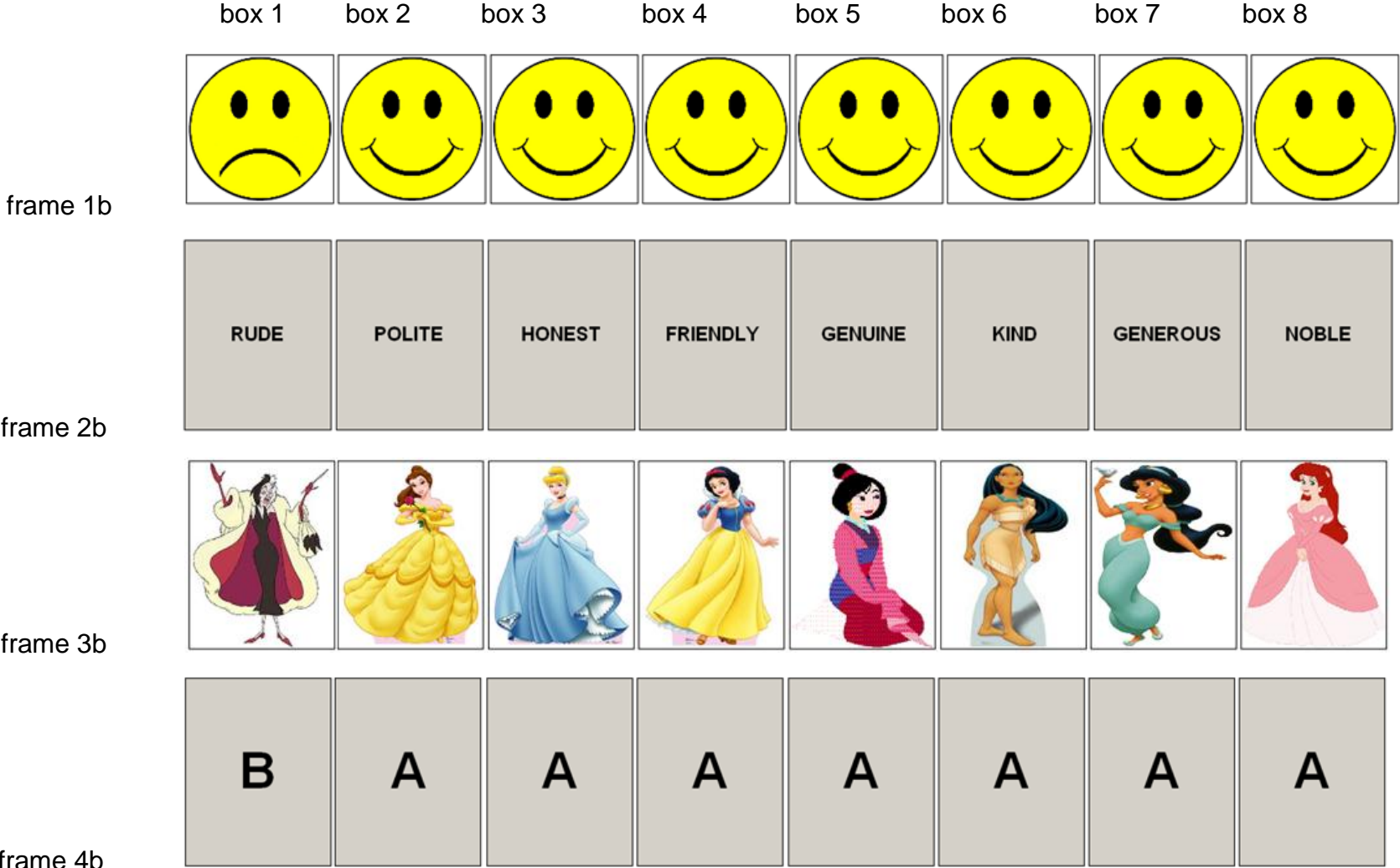
set of boxes	percentage of choices that are of box 1 ( $n = 40$ ):	
	four-box games	eight-box-games
1a, 1b	72.5	85.0
2a, 2b	70.0	57.5
3a, 3b	67.5	60.0
4a, 4b	80.0	82.5
5a, 5b	80.0	95.0
6a, 6b	92.5	92.5
7a, 7b	85.0	87.5
8a, 8b	92.5	87.5
9a, 9b	87.5	92.5
10a, 10b	52.5	40.0
11a, 11b	25.0	27.5
12a, 12b	67.5	70.0
13a, 13b	80.0	80.0
14a, 14b	67.5	62.5
15a, 15b	85.0	92.5
16a, 16b	82.5	90.0
17a, 17b	90.0	80.0
18a, 18b	90.0	85.0
all sets	76.0	76.0

**TABLE V: FREQUENCY OF ODDITY CHOICES BY PLAYERS IN ALL ROLES**

role	percentage of all choices that are of oddity:		
	lower bound	observed	upper bound
<i>four-box games</i>			
coordinators ( $n = 40$ )	68.6	76.0	83.3
discoordinators ( $n = 40$ )	17.6	23.3	29.1
hiders (part 1: $n = 100$ )	12.0	14.6	17.2
hiders (part 2: $n = 100$ )	14.6	18.2	21.9
seekers (part 1: $n = 100$ )	22.1	25.8	29.4
seekers (part 2: $n = 100$ )	17.3	20.8	24.3
<i>eight-box games</i>			
coordinators ( $n = 40$ )	69.6	76.0	82.4
discoordinators ( $n = 40$ )	9.3	15.4	21.5
hiders (part 1: $n = 100$ )	5.4	8.6	11.7
hiders (part 2: $n = 100$ )	6.8	9.8	12.8
seekers (part 1: $n = 100$ )	21.2	25.3	29.5
seekers (part 2: $n = 100$ )	11.9	16.1	20.4

Note:  $n$  denotes the number of subjects who faced games of the relevant type. Each coordinator and discoordinator faced 18 games of that type; each hider and seeker faced 9 games of that type.

**FIGURE I: FRAMES**

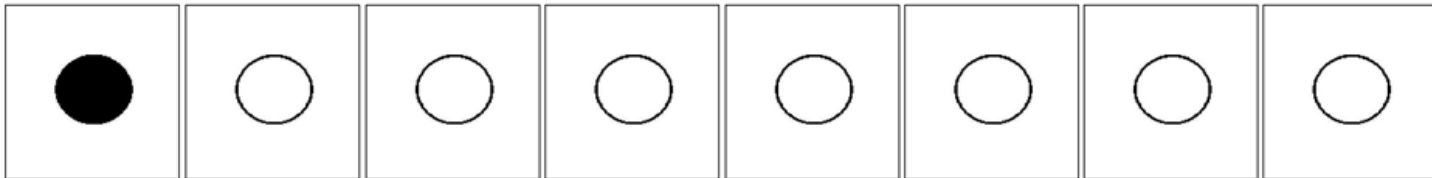


box 1      box 2      box 3      box 4      box 5      box 6      box 7      box 8

frame 5b



frame 6b



frame 7b



frame 8b

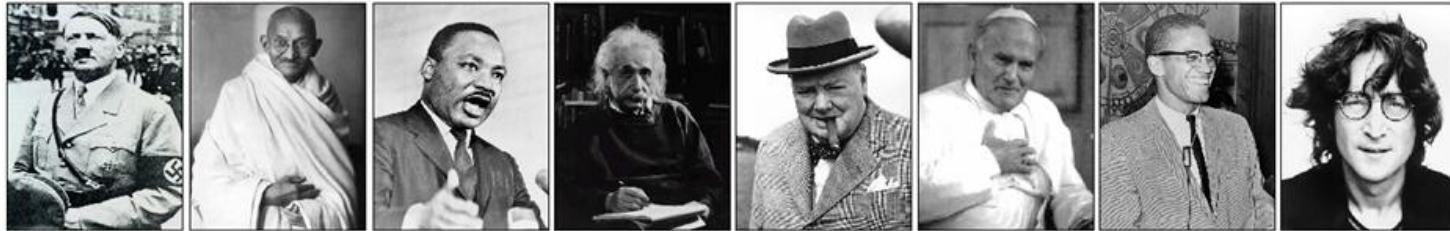


box 1      box 2      box 3      box 4      box 5      box 6      box 7      box 8

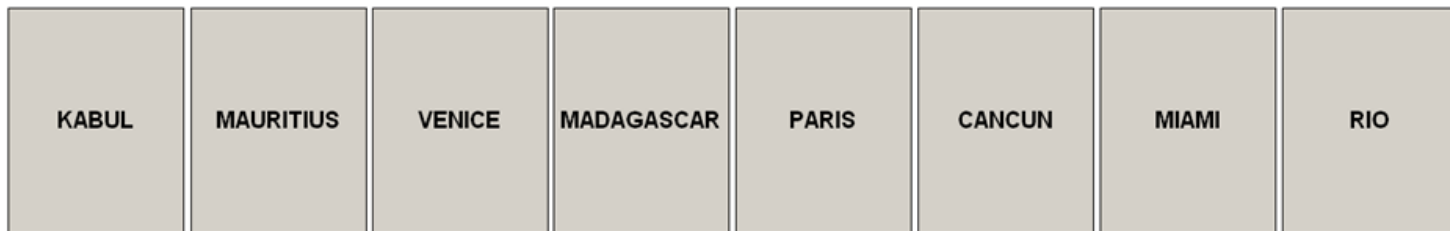
frame 9b



frame 10b









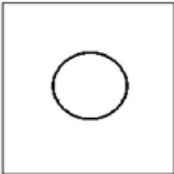
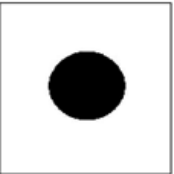
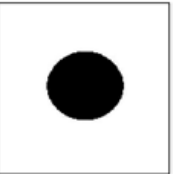
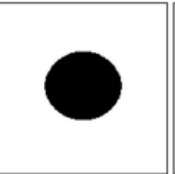
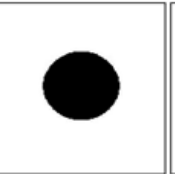
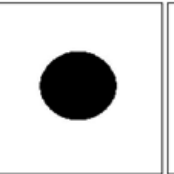
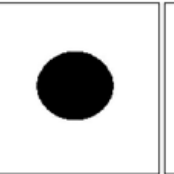
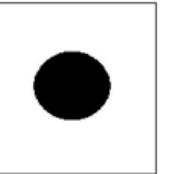










frame 11b



frame 12b



	box 1	box 2	box 3	box 4	box 5	box 6	box 7	box 8
frame 13b	白い	BLUE	RED	BLACK	YELLOW	ORANGE	GREEN	PURPLE
frame 14b								
frame 15b								
frame 16b								

box 1	box 2	box 3	box 4	box 5	box 6	box 7	box 8
<b>FITNESS</b>	<b>CANCER</b>	<b>AIDS</b>	<b>HEPATITIS</b>	<b>CHOLERA</b>	<b>MALARIA</b>	<b>MENINGITIS</b>	<b>PNEUMONIA</b>

frame 17b



frame 18b



**FIGURE II: NUMBERS OF ODDITY CHOICES BY INDIVIDUAL PLAYERS**

