# MODELING BEHAVIOR IN NOVEL STRATEGIC SITUATIONS VIA LEVEL-*K* THINKING

VINCENT P. CRAWFORD
UNIVERSITY OF CALIFORNIA, SAN DIEGO

MARKETING SEMINAR, 3 APRIL 2008
HAAS SCHOOL OF BUSINESS
UNIVERSITY OF CALIFORNIA, BERKELEY

APPLIED MICRO THEORY WORKSHOP, 28 APRIL 2008
UNIVERSITY OF PENNSYLVANIA

GAMES 2008, THIRD WORLD CONGRESS
OF THE GAME THEORY SOCIETY, 14 JULY 2008
NORTHWESTERN UNIVERSITY

# Introduction

Recent experiments suggest that in strategic settings without clear precedents, people's initial responses often deviate systematically from equilibrium.

Experimental evidence suggests that in such settings a structural non-equilibrium model based on "level-*k* thinking"—or a "cognitive hierarchy" model, as Camerer, Ho, and Chong (2004 *QJE*; "CHC") call their closely related model—can often out-predict equilibrium.

The evidence also suggests that level-*k* models can out-predict "equilibrium with noise" models with payoff-sensitive error distributions, such as quantal response equilibrium ("QRE").

The talk begins with an introduction to level-*k* models and a sampling of the supporting experimental evidence.

It then illustrates the use of level-*k* models in several applications involving people's (usually experimental subjects') initial responses to novel strategic situations and the adaptations that are required in them.

The illustrations show that level-*k* models are a flexible, tractable, and useful modeling tool.

They suggest that level-*k* models can often explain more of the variation in initial responses than equilibrium or QRE, and that it can help to resolve empirical puzzles in applications of game theory.

The applications I will discuss include (adaptations in parentheses):

- Crawford and Iriberri's (2007 *AER*) explanation of systematic deviations from unique mixed-strategy equilibrium in O'Neill's (1987 *PNAS*) and Rubinstein, Tversky, and Heller's (1996) zero-sum two-person "hide-and-seek" games (non-neutral framing)
- Crawford, Gneezy, and Rottenstreich's (2008 in press *AER*; "CGR") explanation of coordination and miscoordination in Schelling-style coordination games (non-neutral framing)
- Crawford and Iriberri's (2007 *ECMA*) analysis of systematic overbidding in independent-private-value and common-value auctions (incomplete information)
- Crawford's (2003 *AER*) analysis of deceptive preplay communication of intentions in zero-sum two-person games (extensive-form games)
- Crawford's (2007) analysis of preplay communication of intentions in coordination games as studied by Farrell (1987 *Rand Journal*) and Rabin (1994 *JET*) (extensive-form games)

Other interesting applications include (list not comprehensive):

- CHC's analysis of tacit coordination via structure and "magical" ex post coordination in market-entry games (normal-form games)
- CHC's analysis of speculation and zero-sum betting (incomplete information)
- CHC's analysis of money illusion in coordination (normal-form games)
- Blume et al.'s (2001 *GEB*), Cai and Wang's (2006 *GEB*), Sánchez-Pagés and Vorsatz's (2007 GEB, 2007), Kawagoe and Tazikawa's (2008 *GEB*), and Wang, Spezio, and Camerer's (2006) analyses of "overcommunication" in sender-receiver games (extensive-form games, incomplete information)
- Ellingsen and Östling's (2007) analyses of Aumann's (1990) critique, symmetry-breaking, and reassurance in coordination games with one- and two-sided preplay communication of intentions (extensive-form games)
- Crawford, Kugler, Neeman, and Pauzner's (2008, in progress) analysis of behaviorally optimal (level-*k*) auction design

# Level-*k* models

Level-*k* models were introduced to describe experimental data by Stahl and Wilson (1994 *JEBO*, 1995 *GEB*) and Nagel (1995 *AER*).

Level-*k* models were further studied experimentally by Ho, Camerer, and Weigelt (1998 *AER*; "HCW"); Costa-Gomes et al. (2001 *ECMA*; "CGCB"), Costa-Gomes and Weizsäcker (2008 *RES*), and Costa-Gomes and Crawford (2006 *AER*; "CGC").

Level-*k* models allow behavior to be heterogeneous, but assume that each player follows a rule drawn from a common distribution over a particular hierarchy of decision rules or *types* (as they are called in this literature; no relation to "types" as realizations of private information).

Type *Lk* anchors its beliefs in a nonstrategic *L0* type, which is meant to describe *Lk*'s model of others' instinctive reactions to the game.

*Lk* then adjusts its beliefs via thought-experiments with iterated best responses: *L1* best responds to *L0*, *L2* to *L1*, and so on.

In applications the population type frequencies are treated as behavioral parameters, to be estimated from the data or translated or extrapolated from previous analyses.

The estimated type distribution is typically fairly stable across games, with most weight on *L1*, *L2*, and perhaps *L3*.

The estimated frequency of the anchoring *L0* type is usually small.

Thus, *L0* "exists" mainly as *L1*'s model of others, *L2*'s model of *L1*'s model of others, and so on.

Even so, the specification of *L0* is the main issue in defining a level-*k* model and the key to its explanatory power.

*L0* needs to be adapted to the setting as illustrated below, but the definition of higher types via iterated best responses allows a simple, reliable explanation of behavior across different settings.

Alternative specifications of level-*k* types have been considered:

● Stahl and Wilson have some higher types ("*Worldly*") that best respond to mixtures of noisy versions of lower types.

● CHC have *Lk* best responding to an estimated mixture of lower types, via a one-parameter Poisson type distribution.

My co-authors and I prefer the simpler specification above, which is at least as consistent with the evidence and more tractable.

Estimating an unconstrained type distribution also provides a useful diagnostic: If the data can only be fitted by a weird type distribution—non-hump-shaped (in a homogeneous population) or with implausibly high frequencies of higher types—the explanation is not credible.

*L1* and higher types have accurate models of the game and are decision-theoretically rational, in that they choose best responses to beliefs (in many games, *Lk* makes *k*-rationalizable decisions).

*Lk*'s only departure from equilibrium is in replacing its assumed perfect model of others' decisions with simplified models that avoid the complexity of equilibrium analysis.

Compare Selten (1998 *EER*):

> "Basic concepts in game theory are often circular in the sense that they are based on definitions by implicit properties…. Boundedly rational strategic reasoning seems to avoid circular concepts. It directly results in a procedure by which a problem solution is found."

Although the level-*k* approach, like equilibrium, is a general model of strategic behavior, the two are complements, not competitors.

We all believe that equilibrium (or self-confirming equilibrium, etc.) is a reliable model of people's limiting behavior in situations where they have had enough experience from repeated play in stable settings to learn to predict each others' responses.

But even if eventual convergence to equilibrium is assured, in novel strategic situations we need a reliable model of initial responses.

In such situations a level-*k* analysis can establish the robustness of equilibrium predictions in games where level-*k* types mimic equilibrium.

It can also challenge the conclusions of equilibrium analyses of games where equilibrium is implausible without learning.

And it can resolve empirical puzzles by explaining the systematic deviations from equilibrium such games often evoke.

Level-*k* analyses can also elucidate coordination in situations where learning assures eventual convergence to equilibrium but equilibrium refinements are not a reliable guide to equilibrium selection.

In such situations the limiting equilibrium is jointly determined by the rules that describe the learning process and which of the basins of attraction those rules create people's initial responses fall into.

Consider, for example, the "Continental Divide" coordination game from Van Huyck, Cook, and Battalio's (1997 *JEBO*) experiment, discussed in Camerer (*Behavioral Game Theory*, 2003, Chapter 1).

Seven subjects choose simultaneously and anonymously among "effort" levels from 1 to 14, with each subject's payoff determined by his own effort and a summary statistic, the median, of all players' efforts.

The group median is then publicly announced, subjects choose new effort levels, and the process continues.

# Continental divide game payoffs

| | Median Choice | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| your choice | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| 1 | 45 | 49 | 52 | 55 | 56 | 55 | 46 | -59 | -88 | -105 | -117 | -127 | -135 | -142 |
| 2 | **48** | 53 | 58 | 62 | 65 | 66 | 61 | -27 | -52 | -67 | -77 | -86 | -92 | -98 |
| 3 | **48** | **54** | **60** | **66** | 70 | 74 | 72 | 1 | -20 | -32 | -41 | -48 | -53 | -58 |
| 4 | 43 | 51 | 58 | 65 | **71** | **77** | 80 | 26 | 8 | -2 | -9 | -14 | -19 | -22 |
| 5 | 35 | 44 | 52 | 60 | 69 | **77** | **83** | 46 | 32 | 25 | 19 | 15 | 12 | 10 |
| 6 | 23 | 33 | 42 | 52 | 62 | 72 | 82 | 62 | 53 | 47 | 43 | 41 | 39 | 38 |
| 7 | 7 | 18 | 28 | 40 | 51 | 64 | 78 | 75 | 69 | 66 | 64 | 63 | 62 | 62 |
| 8 | -13 | -1 | 11 | 23 | 37 | 51 | 69 | 83 | 81 | 80 | 80 | 80 | 81 | 82 |
| 9 | -37 | -24 | -11 | 3 | 18 | 35 | 57 | 88 | 89 | 91 | 92 | 94 | 96 | 98 |
| 10 | -65 | -51 | -37 | -21 | -4 | 15 | 40 | **89** | **94** | 98 | 101 | 104 | 107 | 110 |
| 11 | -97 | -82 | -66 | -49 | -31 | -9 | 20 | 85 | **94** | **100** | 105 | 110 | 114 | 119 |
| 12 | -133 | -117 | -100 | -82 | -61 | -37 | -5 | 78 | 91 | 99 | **106** | **112** | **118** | **123** |
| 13 | -173 | -156 | -137 | -118 | -96 | -69 | -33 | 67 | 83 | 94 | 103 | 110 | 117 | **123** |
| 14 | -217 | -198 | -179 | -158 | -134 | -105 | -65 | 52 | 72 | 85 | 95 | 104 | 112 | 120 |

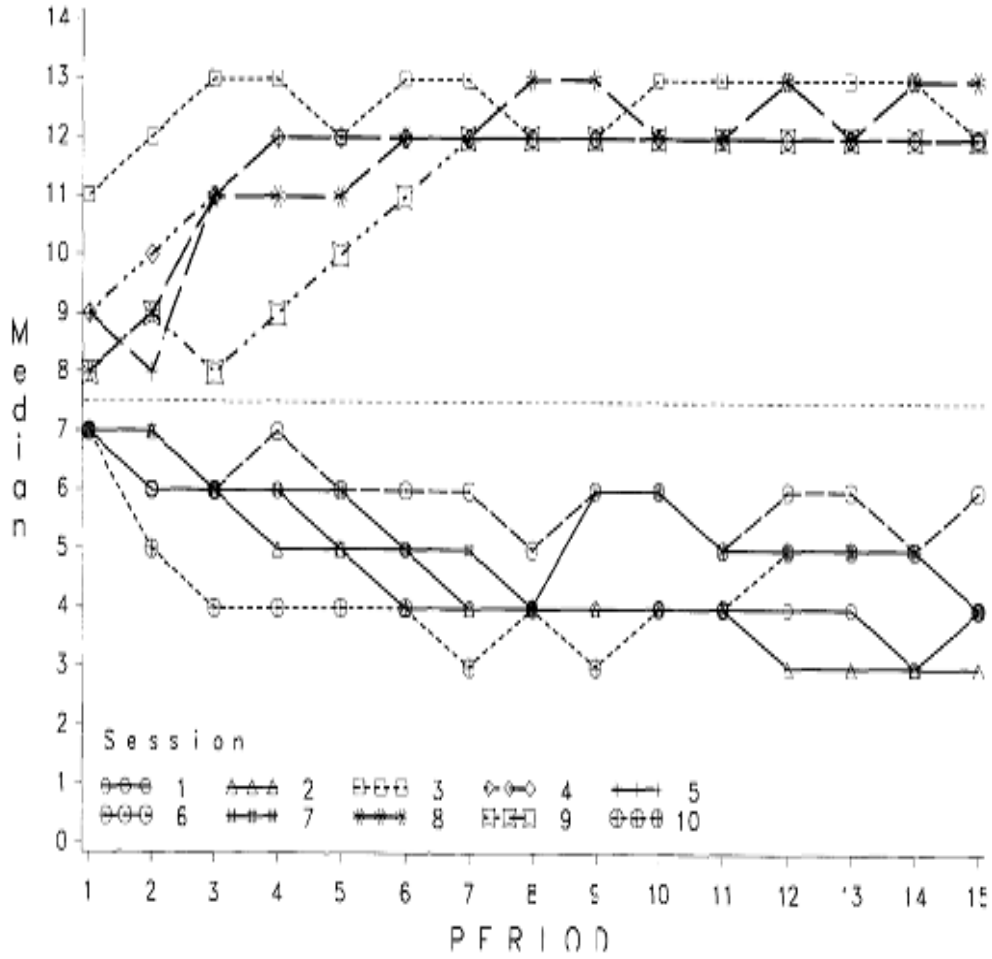



Fig. 3. Median choice in sessions 1 to 10 by period.

There is near-perfect lock-in on the equilibrium determined by which basin of attraction people's initial responses fall into. Without a model of (the prior distribution of) initial responses, prediction is impossible.

# Experimental evidence for level-*k* models

Camerer (*Behavioral Game Theory*, 2003, Chapter 5), CHC (Section IV), and CGC (Introduction, Section II.D) summarize the experimental evidence for level-*k* models in games with a variety of structures.

Here I give the flavor of the evidence by summarizing CGC's results.

CGC's experiments randomly and anonymously paired subjects to play series of 2-person guessing games, with no feedback; the designs suppress learning and repeated-game effects in order to elicit subjects' initial responses, game by game.

The goal was to focus on how players model others' decisions by studying strategic thinking "uncontaminated" by learning.

("Eureka!" learning was possible, but can be tested for and is rare.)

In CGC's guessing games, each player has his own lower and upper limit, both strictly positive (finite dominance-solvability).

Each player also has his own target, and his payoff increases with the closeness of his guess to his target times the other's guess.

The targets and limits vary independently across players and games, with targets both less than one, both greater than one, or "mixed".

(In previous guessing experiments, the targets and limits were always the same for both players, and they varied at most across treatments.)

The games have essentially unique equilibria determined (but not always directly) by players' lower (upper) limits when the product of targets is less (greater) than one.

Consider for example a game (γ2β4 in our notation) in which players' targets are 0.7 and 1.5, the first player's limits are [300, 500], and the second player's limits are [100, 900].

The product of players' targets is 1.05 > 1; in equilibrium the first player guesses his upper limit of 500, but the second player guesses 750, below his upper limit of 900.

When the product of targets is < 1, the equilibrium is determined by the lower limits in a similar way.

The discontinuity of the equilibrium correspondence when the product of targets equals one stress-tests equilibrium, which responds much more strongly to the product of the targets than alternative rules do, and enhances the separation of equilibrium from alternative rules. It also reveals other interesting patterns, not discussed here.

In standard normal-form games, most of the evidence suggests defining *L0* as uniform random over the feasible range of decisions.

In addition to *Equilibrium* and the level-*k* types *L1*, *L2*, and *L3*, CGC's data analysis considered two "iterated dominance" types:

- *D1*, which does one round of dominance and then best responds to a uniform prior over its partner's remaining decisions

- *D2*, which does two rounds and then best responds to a uniform prior over its partner's remaining decisions

- CGC also considered a *Sophisticated* type, which best responds to the probability distributions of others' decisions (estimated from observed frequencies).

*Sophisticated* is the behavioral game theory ideal, included to learn whether any subjects have an understanding of others' decisions that transcends mechanical rules.

CGC's large strategy spaces and the independent variation of targets and limits across games enhance the separation of types' implications, to the point where many subjects' types can be precisely identified from their guessing "fingerprints":

**Types' guesses in the 16 games, in (randomized) order played**

| | *L1* | *L2* | *L3* | *D1* | *D2* | *Eq.* | *Soph.* |
|---|---|---|---|---|---|---|---|
| **1** | 600 | 525 | 630 | 600 | 611.25 | 750 | 630 |
| **2** | 520 | 650 | 650 | 617.5 | 650 | 650 | 650 |
| **3** | 780 | 900 | 900 | 838.5 | 900 | 900 | 900 |
| **4** | 350 | 546 | 318.5 | 451.5 | 423.15 | 300 | 420 |
| **5** | 450 | 315 | 472.5 | 337.5 | 341.25 | 500 | 375 |
| **6** | 350 | 105 | 122.5 | 122.5 | 122.5 | 100 | 122 |
| **7** | 210 | 315 | 220.5 | 227.5 | 227.5 | 350 | 262 |
| **8** | 350 | 420 | 367.5 | 420 | 420 | 500 | 420 |
| **9** | 500 | 500 | 500 | 500 | 500 | 500 | 500 |
| **10** | 350 | 300 | 300 | 300 | 300 | 300 | 300 |
| **11** | 500 | 225 | 375 | 262.5 | 262.5 | 150 | 300 |
| **12** | 780 | 900 | 900 | 838.5 | 900 | 900 | 900 |
| **13** | 780 | 455 | 709.8 | 604.5 | 604.5 | 390 | 695 |
| **14** | 200 | 175 | 150 | 200 | 150 | 150 | 162 |
| **15** | 150 | 175 | 100 | 150 | 100 | 100 | 132 |
| **16** | 150 | 250 | 112.5 | 162.5 | 131.25 | 100 | 187 |

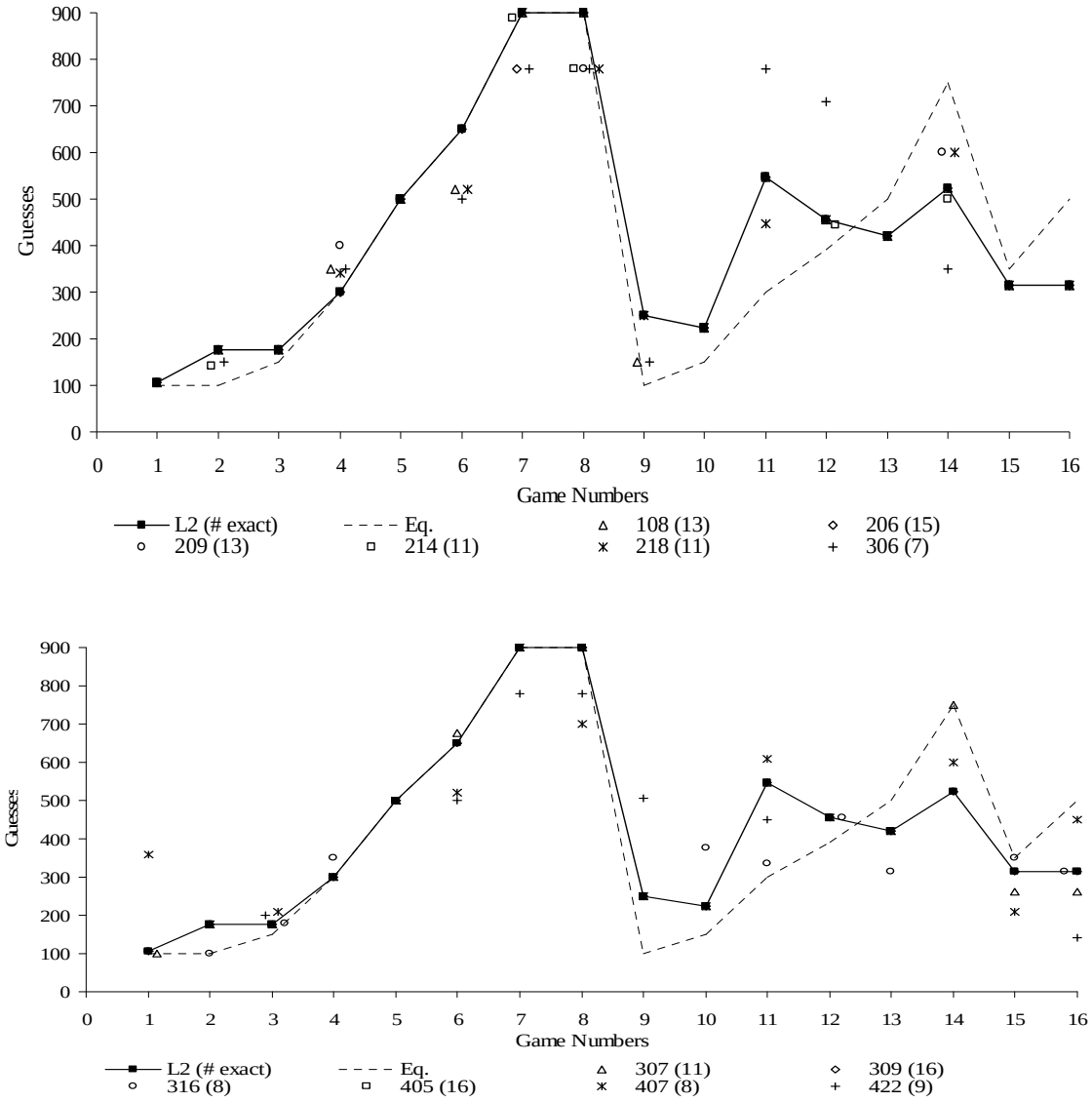
Of the 88 subjects in CGC's main treatments, 43 made guesses that complied *exactly* (within 0.5) with one type's guesses in from 7 to 16 of the games (20 *L1*, 12 *L2*, 3 *L3*, and 8 *Equilibrium*).

For example, CGC's Figure 2 shows the "fingerprints" of the 12 subjects whose guesses conformed most closely to *L2*'s; 72% of these guesses were exact; only the deviations are shown.

The size of CGC's strategy spaces, with 200 to 800 possible exact guesses per game, and the fact that each subject played 16 different games, makes exact compliance very powerful evidence for the type whose guesses are tracked.

If, say, a subject chooses 525, 650, 900, 546 in games 1 to 4, we "know" that he's *L2*.

Further, because CGC's definition of *L2* builds in risk-neutral, self-interested rationality, we also know that the subject's deviations from equilibrium are "caused" not by irrationality, risk aversion, altruism, spite, or confusion, but by his simplified model of others.




**CGC's Figure 2. "Fingerprints" of 12 Apparent *L2* Subjects**

CGC's other 45 subjects made guesses that conformed less closely to a type, but econometric estimates of their types are concentrated on *L1*, *L2*, *L3*, and *Equilibrium*, in roughly the same proportions.

TABLE 1—SUMMARY OF BASELINE AND OB SUBJECTS' ESTIMATED TYPE DISTRIBUTIONS

| Type | Apparent from guesses | Econometric from guesses | Econometric from guesses, excluding random | Econometric from guesses, with specification test | Econometric from guesses and search, with specification test |
|---|---|---|---|---|---|
| *L1* | 20 | 43 | 37 | 27 | 29 |
| *L2* | 12 | 20 | 20 | 17 | 14 |
| *L3* | 3 | 3 | 3 | 1 | 1 |
| *D1* | 0 | 5 | 3 | 1 | 0 |
| *D2* | 0 | 0 | 0 | 0 | 0 |
| *Eq.* | 8 | 14 | 13 | 11 | 10 |
| *Soph.* | 0 | 3 | 2 | 1 | 1 |
| Unclassified | 45 | 0 | 10 | 30 | 33 |

*Note:* The far-right-hand column includes 17 OB subjects classified by their econometric-from-guesses type estimates.

CGC's Table 1

Because *Lk* makes *k*-rationalizable decisions, it is tempting to take the high frequencies of *Lk* guesses as evidence that subjects are explicitly performing finitely iterated dominance; theorists often interpret the spikes in Nagel's (1995 *AER*) data this way.

Although *Dk*'s and *Lk*+1's guesses are perfectly confounded in Nagel's main games, CGC's design strongly separates them.

CGC's data analysis shows that there are essentially no *Dk* types, hence that most of CGC's subjects who respect finitely iterated dominance did so because they were following *Lk* types that mimic iterated dominance, not because they were explicitly performing it.

CGC's analysis also shows that there are no *Sophisticated* types, and that CGC's subjects whose guesses are closest to *Equilibrium* are actually following types that only mimic equilibrium in some games.

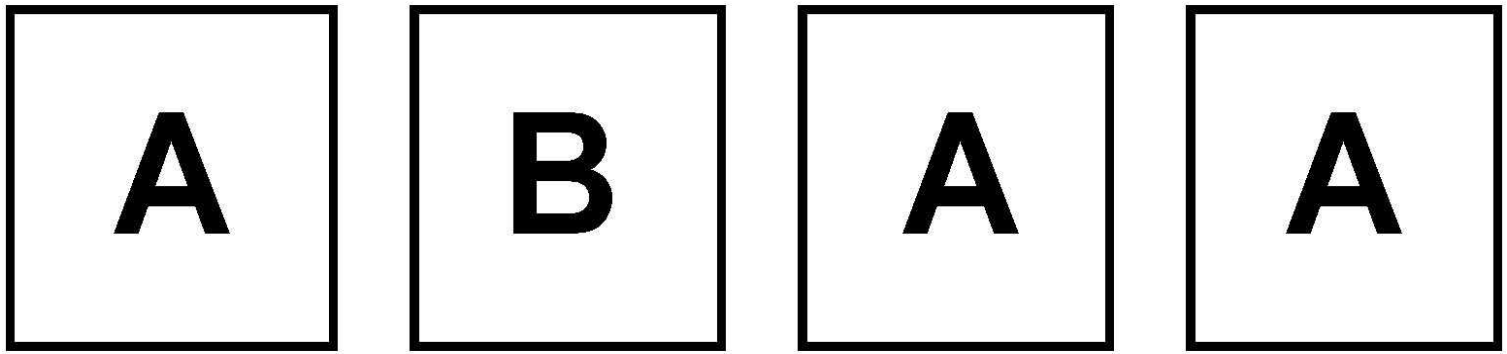
Further, CGC's data strongly resist an "equilibrium plus noise" or QRE interpretation, and subjects' "errors" usually appear to be structural or cognitive, without the payoff-sensitivity a QRE interpretation requires.

## Application: Crawford and Iriberri's (2007 *AER*) explanation of systematic deviations from unique mixed-strategy equilibrium in zero-sum two-person "hide-and-seek" games with non-neutral framing of locations

Consider Rubinstein, Tversky, and Heller's (1993, 1996, 1998-99; "RTH") hide and seek games with non-neutral framing of locations.

A typical seeker's instructions (hider's instructions are analogous):

> *Your opponent has hidden a prize in one of four boxes arranged in a row. The boxes are marked as shown below: A, B, A, A. Your goal is, of course, to find the prize. His goal is that you will not find it. You are allowed to open only one box. Which box are you going to open?*



RTH's framing of the hide and seek game is non-neutral in two ways:

- The "*B*" location is distinguished by its label

- The two "*end A*" locations may be inherently focal

This gives the "*central A*" location its own brand of uniqueness as the "least salient" location.

Mathematically this uniqueness is analogous to the uniqueness of "*B*", but the analysis shows that its psychological effects are quite different.

RTH's design is important as a tractable abstract model of a non-neutral cultural or geographic frame, or "landscape".

Similar landscapes are common in "folk game theory":

- "Any government wanting to kill an opponent…would not try it at a meeting with government officials."
  (comment on the poisoning of Ukrainian presidential candidate—now president—Viktor Yushchenko)

(The meeting with government officials is analogous to RTH's B, but there's nothing in this example analogous to the end locations.)

- "…in Lake Wobegon, the correct answer is usually 'c'."
  (Garrison Keillor (1997) on multiple-choice tests)

(With four possible choices arrayed left to right, this example is very close to RTH's design.)

Perhaps as a result, RTH's design even made it into an episode of the CBS series *Numb3rs*, "Assassin" (clip at http://www.youtube.com/watch?v=HCinK2PUfyk):

Charlie: Hide and seek.

Don: What are you talking about, like the kids' version?

Charlie: A mathematical approach to it, yes. See, the assassin must hide in order to accomplish his goal, we must seek and find the assassin before he achieves that goal.

Megan: Ah, behavioral game theory, yeah, we studied this at Quantico.

Charlie: I doubt you studied it the way that Rubinstein, Tversky and Heller studied two person constant sum hide and seek with unique mixed strategy equilibria.

Megan: No, not quite that way.

Don: Just bear with him.

Hide-and-seek has a clear equilibrium prediction, which leaves no room for framing to systematically influence the outcome.

Yet framing has a strong and systematic effect, qualitatively the same around the world, with *Central A* (or its analogs in other treatments, as explained in the paper) most prevalent for hiders (37% in the aggregate) and even more prevalent for seekers (46%).

TABLE 1—AGGREGATE CHOICE FREQUENCIES IN RTH'S TREATMENTS

| RTH-4 | A | *B* | A | A |
|---|---|---|---|---|
| Hider (53; $p = 0.0026$) | 9 percent | 36 percent | 40 percent | 15 percent |
| Seeker (62; $p = 0.0003$) | 13 percent | 31 percent | 45 percent | 11 percent |
| RT-AABA–Treasure | A | A | *B* | A |
| Hider (189; $p = 0.0096$) | 22 percent | 35 percent | 19 percent | 25 percent |
| Seeker (85; $p = 9E\text{-}07$) | 13 percent | 51 percent | 21 percent | 15 percent |
| RT-AABA–Mine | A | A | *B* | A |
| Hider (132; $p = 0.0012$) | 24 percent | 39 percent | 18 percent | 18 percent |
| Seeker (73; $p = 0.0523$) | 29 percent | 36 percent | 14 percent | 22 percent |
| RT-1234–Treasure | *1* | *2* | 3 | 4 |
| Hider (187; $p = 0.0036$) | 25 percent | 22 percent | 36 percent | 18 percent |
| Seeker (84; $p = 3E-05$) | 20 percent | 18 percent | 48 percent | 14 percent |
| RT-1234–Mine | *1* | *2* | 3 | 4 |
| Hider (133; $p = 6E\text{-}06$) | 18 percent | 20 percent | 44 percent | 17 percent |
| Seeker (72; $p = 0.149$) | 19 percent | 25 percent | 36 percent | 19 percent |
| R-ABAA | A | *B* | A | A |
| Hider (50; $p = 0.0186$) | 16 percent | 18 percent | 44 percent | 22 percent |
| Seeker (64; $p = 9E\text{-}07$) | 16 percent | 19 percent | 54 percent | 11 percent |

*Notes:* Sample sizes and $p$-values for significant differences from equilibrium in parentheses; salient labels in italics; order of presentation of locations to subjects as shown.

Folk game theory deviates from equilibrium logic in similar ways: Any game theorist worth his salt would respond to the Yushchenko quote:

"Any government wanting to kill an opponent…would not try it at a meeting with government officials."

with

"If investigators thought that way, a meeting with government officials is precisely where a government *would* try to kill an opponent."

**Puzzles:**

- Hiders' and seekers' responses are unlikely to be completely non-strategic in such simple games. So if they aren't following equilibrium logic, what are they doing?

- On average hiders are as smart as seekers, so hiders tempted to hide in *central A* should realize that seekers will be just as tempted to look there. So, why do hiders allow seekers to find them 32% of the time when they could hold it down to 25% via the equilibrium mixed strategy?

- Further, why do seekers choose *central A* even more often than hiders? Although the payoff structure of RTH's game is asymmetric, QRE coincides with equilibrium in them, and so does not help to explain this asymmetry of choice distributions.

**Resolution:**

QRE ignores labeling and (despite the game's payoff asymmetry) coincides with equilibrium, and so cannot help explain the deviations.

The role asymmetry in behavior strongly suggests something like a level-*k* explanation (and is a mystery from the viewpoint any other theory we are aware of).

Here defining *L0* as uniform random would be unnatural given the non-neutral framing of decisions and that *L0* describes others' instinctive responses. (It would also make *Lk* the same as *Equilibrium* for $k > 0$.)

But a level-*k* model with a role-independent *L0* that probabilistically favors salient locations yields a simple explanation of RTH's results.

We assume that *L0* hiders and seekers both choose A, B, A, A with probabilities $p/2$, $q$, $1 - p - q$, $p/2$ respectively, with $p > ½$ and $q > ¼$.

*L0* favors both the end locations and the B location, equally for hiders and seekers, but we leave it open which is more salient.

A level-*k* model gracefully explains the main patterns in RTH's data:

- Given *L0*'s attraction to salient locations, *L1* hiders choose *central A* to avoid *L0* seekers and *L1* seekers avoid *central A* in searching for *L0* hiders (the data suggest that end locations are more salient than B)
- For similar reasons, *L2* hiders choose *central A* with probability between 0 and 1 and *L2* seekers choose it with probability 1
- *L3* hiders avoid *central A* and *L3* seekers choose it with probability between zero and one
- *L4* hiders and seekers both avoid *central A*

For plausible type distributions (estimated 19% *L1*, 32% *L2*, 24% *L3*, 25% *L4*—almost hump-shaped), the model explains the prevalence of *central A* for hiders and its even greater prevalence for seekers.

The role asymmetry in behavior follows naturally from hiders' and seekers' asymmetric responses to *L0*'s *role-symmetric* choices. (However, only a heterogeneous population with substantial frequencies of *L2* and *L3* as well as *L1* can reproduce the patterns.)

The analysis suggests that our first epigraph ("Any government wanting to kill an opponent…would not try it at a meeting with government officials") reflects the reasoning of an *L1* poisoner, or equivalently of an *L2* investigator reasoning about an *L1* poisoner.

RTH took the main patterns in their data as evidence that their subjects did not think strategically:

● "The finding that both choosers and guessers selected the least salient alternative suggests little or no strategic thinking."

● "In the competitive games, however, the players employed a naïve strategy (avoiding the endpoints), that is not guided by valid strategic reasoning. In particular, the hiders in this experiment either did not expect that the seekers too, will tend to avoid the endpoints, or else did not appreciate the strategic consequences of this expectation."

But our analysis suggests that their subjects were actually quite strategic and in fact unusually sophisticated (with a substantial fraction of *L3*s and even some *L4*s)—they just didn't follow equilibrium logic.

## Aside on model evaluation

Although our empirically based prior about the hump shape and location of the type distribution imposes some discipline, the freedom to specify *L0* leaves room for doubt about overfitting and portability.

To see if our proposed level-*k* explanation is more than a "just-so" story, we compare it on the overfitting and portability dimensions with the leading alternatives:

- Equilibrium with intuitive payoff perturbations (salience lowers hiders' payoffs, other things equal; while salience raises seekers' payoffs)
- QRE with similar payoff perturbations
- Alternative level-*k* specifications

We test for overfitting by re-estimating each model separately for each of RTH's six treatments and using the re-estimated models to "predict" the choice frequencies of the other treatments.

Our favored level-*k* model has a modest prediction advantage over the alternative models, with mean squared prediction error 18% lower and better predictions in 20 of 30 comparisons.

A more challenging test regards portability, the extent to which a model estimated from subjects' responses to one game can be extended to predict or explain other subjects' responses to different games.

We consider the two closest relatives of RTH's games in the literature:

- O'Neill's (1987 *PNAS*) famous card-matching game
- Rapoport and Boebel's (1992 *GEB*) closely related game

These games both raise the same kinds of strategic issues as RTH's games, but with more complex patterns of wins and losses, different framing, and in the latter case five locations.

We test for portability by using the leading alternative models, estimated from RTH's data, to "predict" subjects' initial responses in O'Neill's and Rapoport and Boebel's games.

In O'Neill's game, for example, players simultaneously and independently choose one of four cards: A, 2, 3, J.

One player, say the row player (the game was presented to subjects as a story, not a matrix) wins if there is a match on J or a mismatch on A, 2, or 3; the other player wins in the other cases.

| | **A (s)** | **2 (s)** | **3 (s)** | **J (h)** |
|---|---|---|---|---|
| **A (h)** | 1<br>0 | 0<br>1 | 0<br>1 | 1<br>0 |
| **2 (h)** | 0<br>1 | 1<br>0 | 0<br>1 | 1<br>0 |
| **3 (h)** | 0<br>1 | 0<br>1 | 1<br>0 | 1<br>0 |
| **J (s)** | 1<br>0 | 1<br>0 | 1<br>0 | 0<br>1 |

**O'Neill's Card-Matching Game**

O’Neill’s game is like a hide-and-seek game, except that each player is a hider (h) for some locations and a seeker (s) for others. Even so, it is clear how to adapt *L0* or payoff perturbations to the game.

A, 2, and 3 are strategically symmetric, and equilibrium (without perturbations) has Pr{A} = Pr{2} = Pr{3} = 0.2, Pr{J} = 0.4.

Discussions of O’Neill’s data have been dominated by an “Ace effect,” whereby when the data are aggregated over all 105 rounds, row and column players respectively played A 22.0% and 22.6% of the time. (O’Neill speculated that “players were attracted by the powerful connotations of an Ace”.)

But it’s difficult (impossible?) to find a behaviorally plausible level-*k* model in which row players play A more than the equilibrium 20%.

(This requires some algebra to see, starting with types’ predicted decisions; see tables A3 and A4 in the paper’s web appendix.)

Fortunately, for initial responses it turns out that there is no Ace effect. Instead there is a Joker effect, a full order of magnitude stronger:

- 8% A, 24% 2, 12% 3, 56% J for rows

- 16% A, 12% 2, 8% 3, 64% J for columns

Unlike an Ace effect, these frequencies *can* be gracefully explained by a level-*k* model in which *L0* probabilistically favors the salient A and J cards. (J's unique payoff role may make it even more salient than A.)

Our analysis suggests that the Ace effect in the aggregated data is due to learning, not salience; if anything is salient, it's the Joker.

The analysis also traces the superior portability of the level-*k* model to the fact that *L0* reflects decision-theoretic rather than strategic considerations, for which the evidence cuts across strategic structures.

(If *L0* were strategic, it would interact with the differences in strategic structure across games in complex ways that stymie generalization.)

## Application: Crawford, Gneezy, and Rottenstreich's (2008 *AER*) explanation of miscoordination in Schelling-style coordination games with non-neutral framing of decisions

CGR randomly paired subjects to play games with non-neutral framing of decisions like those in Schelling's (1960) classic "meeting in NYC" experiments, but (except for a symmetric game like Schelling's games) with payoffs like Battle of the Sexes.

As in Schelling's experiments, there was a commonly observable labeling of decisions:

In unpaid pilots, run in Chicago, the labeling pitted the world-famous Sears Tower versus the little-known AT&T Building across the street.

| | | P2 (90% Sears) | |
|---|---|---|---|
| | | Sears | AT&T |
| P1 (90% Sears) | Sears | 100,100 | 0,0 |
| | AT&T | 0,0 | 100,100 |

**Symmetric**

| | | P2 (58% Sears) | |
|---|---|---|---|
| | | Sears | AT&T |
| P1 (61% Sears) | Sears | 100,101 | 0,0 |
| | AT&T | 0,0 | 101,100 |

**Slight Asymmetry**

| | | P2 (47% Sears) | |
|---|---|---|---|
| | | Sears | AT&T |
| P1 (50% Sears) | Sears | 100,110 | 0,0 |
| | AT&T | 0,0 | 110,100 |

**Moderate Asymmetry**

## Chicago Skyscrapers

**Sears Tower with the AT&T Building in the background on its left**

More formal, paid treatments pitted the abstract label X against Y, with X presumed (and shown) to be more salient than Y.

| | | P2 X | P2 Y |
|---|---|---|---|
| P1 | X | 5,5 | 0,0 |
| | Y | 0,0 | 5,5 |

**Symmetric**

| | | P2 X | P2 Y |
|---|---|---|---|
| P1 | X | 5,5.1 | 0,0 |
| | Y | 0,0 | 5.1,5 |

**Slight Asymmetry**

| | | P2 X | P2 Y |
|---|---|---|---|
| P1 | X | 5,6 | 0,0 |
| | Y | 0,0 | 6,5 |

**Moderate Asymmetry**

| | | P2 X | P2 Y |
|---|---|---|---|
| P1 | X | 5,10 | 0,0 |
| | Y | 0,0 | 10,5 |

**Large Asymmetry**

Like the salience of Sears Tower, the salience of X makes it easy and in principle obvious for subjects to coordinate on the “both-X” equilibrium; and they do this in the symmetric version of the game.

Since Schelling’s experiments with symmetric games, people have assumed that slight payoff asymmetry would not interfere with this.

But even with slight payoff asymmetry, the game poses a new strategic problem because both-X is one player’s favorite way to coordinate but not the other’s.

Just as in a society of men and women playing Battle of the Sexes, in which Ballet is more salient than Fights, there is a tension between the “label salience” of X and the “payoff-salience” of a player’s favorite way to coordinate: Payoff salience reinforces label salience in one player role (P2s) but opposes it for players in the other (P1s).

This tension may lead players to respond asymmetrically, which in this game is bad for coordination.

In CGR's experiments, even slight payoff asymmetries had a large and surprising effect. Here are the observed choice frequencies.

| | | P2 (76% X) | |
|---|---|---|---|
| | | **X** | **Y** |
| **P1 (76% X)** | **X** | **5,5** | **0,0** |
| | **Y** | **0,0** | **5,5** |

**Symmetric**

| | | P2 (28% X) | |
|---|---|---|---|
| | | **X** | **Y** |
| **P1 (78% X)** | **X** | **5,5.1** | **0,0** |
| | **Y** | **0,0** | **5.1,5** |

**Slight Asymmetry**

| | | P2 (61% X) | |
|---|---|---|---|
| | | **X** | **Y** |
| **P1 (33% X)** | **X** | **5,6** | **0,0** |
| | **Y** | **0,0** | **6,5** |

**Moderate Asymmetry**

| | | P2 (60% X) | |
|---|---|---|---|
| | | **X** | **Y** |
| **P1 (36% X)** | **X** | **5,10** | **0,0** |
| | **Y** | **0,0** | **10,5** |

**Large Asymmetry**

Even tiny payoff asymmetries cause a large drop in the expected coordination rate, from 64% (0.64 = 0.76×0.76 + 0.24×0.24) in the symmetric game to 38%, 46%, and 47% in the asymmetric games.

Perhaps more surprisingly (and unlike in the unpaid Chicago Skyscrapers treatment), the pattern of miscoordination reversed as asymmetric games progressed from small to large payoff differences:

- With slightly asymmetric payoffs, most subjects in both roles favored their partners' payoff-salient decisions.
- But with moderate or large asymmetries, most subjects in both roles switched to favoring their own payoff-salient decisions.

**Puzzles:**

- Why didn't subjects in the asymmetric games ignore the payoff asymmetry, which cannot be used to break the symmetry as required for coordination, and use the salience of Sears Tower to coordinate?
- Why did the pattern of miscoordination reverse as the asymmetric games progressed from small to large payoff differences?

**Resolution:**

Standard notions such as QRE ignore labeling, and so cannot help.

A level-k model can gracefully explain the patterns in the data, but again it's important to have an *L0* that realistically describes people's beliefs about others' instinctive reactions to the tension between label- and payoff- salience that seems to drive the results.

CGR assume that *L0* is the same in both player roles, and that it responds instinctively to both label and payoff salience; but with a "payoffs bias" that favors payoff over label salience, other things equal:

- In symmetric games *L0* chooses X with some probability greater than ½.
- In any asymmetric game, (for simplicity only) whether or not label-salience opposes payoff-salience, *L0* chooses its payoff-salient decision with probability $p > ½$.

(These assumptions are consistent with Crawford and Iriberri's *L0* assumptions, because their games had no payoff-salience.)

Under these assumptions about *L0*, *L1*'s and *L2*'s choices for P1 and P2 are completely determined by *p*, the extent of *L0*'s payoff bias.

Except in symmetric games, even though *L0*'s choice probabilities are the same for P1s and P2s, they imply *L1* and *L2* choice probabilities that differ across player roles due to the asymmetric relationships between label and payoff salience for P1s and P2s.

Simple calculations (CGR's Table 3, reproduced next slide) show that a level-*k* model can track the reversal of the pattern of miscoordination between the slightly asymmetric game and the games with moderate or large payoff asymmetries if (and only if) 0.505 (= 5.1/[5.1+5]) < *p* < 0.545 (= 6/[6+5]), so that *L0* has only a modest payoff bias.

If *p* falls into this range and the population frequency of *L1* is 0.7 and that of *L2* is 0.3, close to most previous estimates, the model's predicted choice frequencies differ from the observed frequencies by more than 10% only in the symmetric game, where the model somewhat overstates the homogeneity of the subject pool (Table 3).

| | Symmetric Labeled (SL) | Asymmetric Slight Labeled (ASL) | Asymmetric Moderate Labeled (AML) | Asymmetric Large Labeled (ALL) |
|---|---|---|---|---|
| Payoffs for coordinating on "*X*" | $5, $5 | $5, $5.10 | $5, $6 | $5, $10 |
| Payoffs for coordinating on "*Y*" | $5, $5 | $5.10, $5 | $6, $5 | $10, $5 |
| Pr{X} for P1 *L0* | > ½ | 1-*p* | 1-*p* | 1-*p* |
| Pr{X} for P2 *L0* | > ½ | *p* | *p* | *p* |
| Pr{X} for P1 *L1* | 1 | 1 | 0 | 0 |
| Pr{X} for P1 *L2* | 1 | 0 | 1 | 1 |
| Pr{X} for P2 *L1* | 1 | 0 | 1 | 1 |
| Pr{X} for P2 *L2* | 1 | 1 | 0 | 0 |
| Total P1 predicted Fr{X} | 100% | 100*q*% | 100(1-*q*)% | 100(1-*q*)% |
| Total P1 predicted Fr{X}\|*q*=0.7 | 100% | 70% | 30% | 30% |
| Total P1 observed Fr{X} | 76% | 78% | 33% | 36% |
| Total P2 predicted Fr{X} | 100% | 100(1-*q*)% | 100*q*% | 100*q*% |
| Total P2 predicted Fr{X}\|*q*=0.7 | 100% | 30% | 70% | 70% |
| Total P2 observed Fr{X} | 76% | 28% | 61% | 60% |

**Table 3. *L1*'s and *L2*'s choice probabilities in *X*-*Y* treatments when 0.505 < *p* < 0.545**

The details are as follows:

| | | P2 (76%) *X* | P2 (76%) *Y* |
|---|---|---|---|
| P1 (76%) | *X* | 5,5 | 0,0 |
| | *Y* | 0,0 | 5,5 |

**Symmetric**

● In the symmetric game, with no payoff salience, *L0* favors the salience of X.

● *L1* P1s and P2s therefore both choose X.

● *L2* P1s and P2s do the same.

In this case the model predicts that 100% of P1s and P2s will choose X. Thus, here it makes the same prediction as equilibrium selection based on salience as in a Schelling focal point. This is fairly accurate, but it overstates the homogeneity of the subject pool.

| | | P2 (28%) | |
|---|---|---|---|
| | | ***X*** | ***Y*** |
| **P1 (78%)** | ***X*** | **5,5.1** | **0,0** |
| | ***Y*** | **0,0** | **5.1,5** |

**Slight Asymmetry**

● In the slightly asymmetric game, with $p > 0.505$ (= 5.1/[5.1+5]), the payoff differences are small enough that *L1* P1s choose P2s' payoff-salient decision, X, because *L1* P1s think it is sufficiently likely that *L0* P2s will choose X that X yields them higher expected payoffs.

● *L2* P2s, who best respond to *L1* P1s, thus choose X as well.

● With $p > 0.505$, *L1* P2s choose P1s' payoff-salient decision, Y, because *L1* P2s think it is sufficiently likely that *L0* P1s will choose Y.

● *L2* P1s thus choose Y.

In this case the model predicts that *L1* P1s choose X and *L2* P1s choose Y, while *L1* P2s choose Y and *L2* P2s choose X. Thus, when $q = 0.7$, the model predicts that 70% of P1s will choose X but only 30% of P2s will choose X. This comes reasonably close to the observed frequencies of 78% and 28%.

| | | P2 (61%) | |
|---|---|---|---|
| | | ***X*** | ***Y*** |
| **P1 (33%)** | ***X*** | **5,6** | **0,0** |
| | ***Y*** | **0,0** | **6,5** |

**Moderate Asymmetry**

| | | P2 (60%) | |
|---|---|---|---|
| | | ***X*** | ***Y*** |
| **P1 (36%)** | ***X*** | **5,10** | **0,0** |
| | ***Y*** | **0,0** | **10,5** |

**Large Asymmetry**

- In the games with moderate or large payoff asymmetries, *L0*'s payoffs bias is strong enough, but not too strong ($p < 0.545$ (= 6/[6+5])), that *L1* P1s and P2s both choose their own instead of their partners' payoff- salient decisions, Y for P1s and X for P2s.
- *L2* P1s choose X and *L2* P2s choose Y.

In this case the model predicts that *L1* P1s choose Y and *L2* P1s choose X, while *L1* P2s choose X and *L2* P2s choose Y. Thus, when $q = 0.7$, the model predicts that 30% of P1s will choose X but 70% of P2s will choose X. This is again reasonably close to the observed frequencies of 33-36% and 61-60%.

# Application: Crawford and Iriberri's (2007 *ECMA*) analysis of systematic overbidding in independent-private-value and common-value auctions

| Equilibrium predictions | | |
|---|---|---|
| | **First-Price** | **Second-Price** |
| **Independent-Private-Value Auctions** | **Shaded Bidding** | **Truthful Bidding** |
| **Common-Value Auctions** | **Value Adjustment + Shaded Bidding** | **Value Adjustment** |

**Puzzle:** Systematic overbidding (relative to equilibrium) has been observed in subjects' initial responses to all kinds of auctions (Goeree, Holt, and Palfrey (2002 *JET*), Kagel and Levin (1986 *AER*, 2000), Avery and Kagel (1997 *JEMS*), Garvin and Kagel (1994 *JEBO*)).

(With independent private values, most of the examples that have been studied experimentally do not separate level-*k* from equilibrium bidding strategies, hence our choice to study GHP's results.)

But the literature has proposed completely different explanations of overbidding for private- and common-value auctions:

- Risk-aversion and/or joy of winning for private-value auctions
- Winner's curse for common-value auctions

**Resolution:**

We propose a level-*k* analysis that provides a unified explanation of these results, without invoking risk-aversion and/or joy of winning.

Our analysis extends Kagel and Levin's (1986 *AER*) and Holt and Sherman's (1994 *AER*) analyses of "naïve bidding".

It also builds on Eyster and Rabin's (2005 *ECMA*; "ER") analysis of "cursed equilibrium" and CHC's (2004, Section VI) level-k/cognitive hierarchy analysis of zero-sum betting.

The analysis allows us to explore how to extend level-*k* models to an important class of incomplete-information games.

It also links experiments on auctions to experiments on strategic thinking.

It also allows us to explore the robustness of equilibrium auction theory to failures of the equilibrium assumption.

The key issue is how to specify *L0*; there are two natural possibilities:

- *Random L0* bids uniformly on the interval between the lowest and highest possible values (even if above own realized value)

- *Truthful L0* bids its expected value conditional on its own signal (meaningful here, but not in all incomplete-information games)

In judging these, bear in mind that they describe only the instinctive starting point of a subject's strategic thinking about others.

We have found it best to make *L0* as dumb as possible, letting higher *Lk*s model strategic thinking.

The model constructs separate type hierarchies on these *L0*s, and allows each subject to be one of the types, from either hierarchy.

(*Random* (*Truthful*) *Lk* is *Lk* defined by iterating best responses from *Random* (*Truthful*) *L0*; and is not itself random or truthful).

Given a specification of *L0*, the optimal bid must take into account:

- Value adjustment for the information revealed by winning (only in common-value auctions)
- The bidding trade-off between the higher price paid if the bidder wins and the probability of winning (only in first-price auctions)

With regard to value adjustment, Random *L1* does not condition on winning because Random *L0* bidders bid randomly, hence independently of their values; Random *L1* is "fully cursed" (ER's term).

All other types do condition on winning, in various ways, but this conditioning tends to make bidders' bids strategic substitutes, in that the higher others' bids are, the greater the (negative) adjustment.

Thus, to the extent that Random *L1* overbids, Random *L2* tends to underbid (relative to equilibrium): if it's bad news that you beat equilibrium bidders, it's even worse news that you beat overbidders.

The bidding tradeoff, by contrast, can go either way.

The question, empirically, is whether the distribution of heterogeneous types' bids (for example, a mixture of Random *L1* overbidding and Random *L2* underbidding) fits the data better than the alternatives.

In three of the four leading cases we study, a level-*k* model has an advantage over equilibrium, cursed equilibrium, and/or QRE.

For the remaining case (Kagel and Levin's first-price auction), the most flexible specification of cursed equilibrium has a small advantage.

Except in Kagel and Levin's second-price auctions, the estimated type frequencies are similar to those found in other experiments:

Random and Truthful *L0* have low or zero estimated frequencies, and the most common types are (in order of importance) Random *L1*, Truthful *L1*, Random *L2*, and sometimes *Equilibrium* or Truthful *L2*.

# Application: Crawford's (2003 *AER*) analysis of preplay communication of intentions in zero-sum two-person games

Consider a simple perturbed matching pennies game, viewed as a model of the Allies' choice of where to invade Europe on D-Day:

| | | **Germans** | |
|---|---|---|---|
| | | **Defend Calais** | **Defend Normandy** |
| **Allies** | **Attack Calais** | 1<br>-1 | -2<br>2 |
| | **Attack Normandy** | -1<br>1 | 1<br>-1 |

- Attacking an undefended Calais is better for the Allies than attacking an undefended Normandy, so better for them on average

- Defending an unattacked Normandy is worse for the Germans than defending an unattacked Calais, and so worse for them on average

Now imagine that D-Day is preceded by a message from the Allies to the Germans regarding their intentions about where to attack.

Imagine that the message is (approximately!) cheap talk.

**An Inflatable “Tank” from Operation Fortitude**

In an equilibrium analysis of a zero-sum game preceded by a cheap-talk message regarding intentions, the sender must make his message uninformative, and the receiver must ignore it. Thus the underlying game must be played according to its mixed-strategy equilibrium, and communication can have no effect.

Yet intuition suggests that in many such situations:

- The sender's message and action are part of a single, integrated strategy
- The sender tries to anticipate which message will fool the receiver and chooses it nonrandomly
- The sender's action differs from what he would have chosen with no opportunity to send a message

Moreover, in my stylized version of D-Day:

- The deception succeeded (the Allies faked preparations for invasion at Calais, the Germans defended Calais and left Normandy lightly defended, and the Allies then invaded Normandy)
- But the sender won in the less beneficial of the two possible ways

Admittedly, D-Day is only one datapoint (if that)….

But there's an ancient Chinese antecedent of D-Day, Huarongdao, in which General Cao Cao chooses between two roads, trying to avoid capture by General Kongming (thanks to Duozhe Li of CUHK for the reference to Luo Guanzhong's historical novel, *Three Kingdoms*).

**Huarongdao**

| | | **Kongming** | |
|---|---|---|---|
| | | **Main Road** | **Huarong** |
| **Cao Cao** | **Main Road** | 3<br>-1 | 0<br>1 |
| | **Huarong** | 1<br>0 | 2<br>-2 |

- Cao Cao loses 2 and Kongming gains 2 if Cao Cao is captured
- Both Cao Cao and Kongming gain 1 by taking the Main Road, whether or not Cao Cao is captured—it's important to be comfortable, even if (especially if?) if you think you're about to die

In Huarongdao, essentially the same thing happened as in D-Day: Kongming lit campfires on the Huarong road; Cao Cao was fooled by this into thinking Kongming would wait for him on the *Main Road*; and Kongming captured Cao Cao but only by taking the bad Huarong road.

(The ending however was happy: Kongming later let Cao Cao go.)

In what sense did the “essentially the same thing” happen?

In D-Day the message was literally deceptive but the Germans were fooled because they “believed” it (either because they were credulous or because they inverted the message one too many times).

Kongming's message was literally truthful—he lit fires on the Huarong Road and ambushed Cao Cao there—but Cao Cao was fooled because he inverted the message.

Although the sender’s and receiver’s message strategies and beliefs were different, the end result—what happened in the underlying game—was the same: The sender won, but in the less beneficial way.

Why was Cao Cao fooled by Kongming’s message?

One advantage of using fiction as data (aside from not needing human subjects approval) is that it can reveal cognition (without eye-tracking):

● *Three Kingdoms* gives Kongming’s rationale for sending a deceptively truthful message: “Have you forgotten the tactic of ‘letting weak points look weak and strong points look strong’?”

● It also gives Cao Cao's rationale for inverting Kongming’s message: “Don’t you know what the military texts say? ‘A show of force is best where you are weak. Where strong, feign weakness.’ ”

Why was Cao Cao fooled by Kongming's message?

One advantage of using fiction as data (aside from not needing human subjects approval) is that it can reveal cognition (without eye-tracking):

- *Three Kingdoms* gives Kongming's rationale for sending a deceptively truthful message: "Have you forgotten the tactic of 'letting weak points look weak and strong points look strong'?"
- It also gives Cao Cao's rationale for inverting Kongming's message: "Don't you know what the military texts say? 'A show of force is best where you are weak. Where strong, feign weakness.' "

Cao Cao must have bought a used, out-of-date edition….

As we will see, with *L0* suitably adapted to this setting Cao Cao's rationale resembles *L1* thinking; but Kongming's rationale resembles *L2* thinking.

**Puzzle:**

We can now restate the puzzle more concretely, for both examples:

● Why did the receiver allow himself to be fooled by a costless (hence easily faked) message from an *enemy*?

● If the sender expected his message to fool the receiver, why didn't he reverse it and fool the receiver in the way that would have allowed him to win in the *more* beneficial way? (Why didn't the Allies feint at Normandy and attack at Calais? Why didn't Kongming light fires and ambush Cao Cao on the main road?)

● Was it a coincidence that the same thing happened in both cases?

**Resolution:**

A level-*k* analysis suggests that it was more than a coincidence.

Assume that Allies' and Germans' types are drawn from separate distributions, including both level-*k*, or *Mortal*, types and a fully strategically rational, or *Sophisticated*, type (interesting but rare).

*Mortal* types use step-by-step procedures that generically determine unique, pure strategies, and avoid simultaneous determination of the kind used to define equilibrium (recall the Selten (1998 *EER*) quote).

*Sophisticated* types know everything about the game, including the distribution of *Mortal* types; and play equilibrium in a "reduced game" between *Sophisticated* players, taking *Mortals*' choices as given.

How should *L0* be adapted to an extensive-form game with communication?

Here a uniform random *L0* does not seem natural, at least for senders. Instead *Mortal* types' behaviors regarding the message are anchored on *L0*s based on truthfulness for senders and credulity for receivers, just as in the informal literature on deception.

(The literature has not yet converged on whether *L0* receivers should be defined as credulous or uniform random—compare Ellingsen and Östling (2007)—but the distinction is partly semantic because *L1* receivers' best responses to truthful *L0* senders are credulous.)

*L1* or higher *Mortal* Allied types always expect to fool the Germans, either by lying (like the Allies) or by telling the truth (like Kongming); given this, all such Allied types send a message they expect to make the Germans think they will attack Normandy; and then attack Calais.

If we knew the Allies and Germans were *Mortal*, we could now derive the model's implications from an estimate of type frequencies.

But the analysis can usefully be extended to allow the possibility of *Sophisticated* Allies and Germans.

To do this, note that *Mortals*' strategies are determined independently of each other's and *Sophisticated* players' strategies, and so can be treated as exogenous (even though they affect others' payoffs).

Next, plug in the distributions of *Mortal* Allies' and Germans' independently determined behavior to obtain a "reduced game" between *Sophisticated* Allies and *Sophisticated* Germans.

Because *Sophisticated* players' payoffs are influenced by *Mortal* players' decisions, the reduced game is no longer zero-sum, its messages are not cheap talk, and it has incomplete information.

(The sender's message, which is ostensibly about his intentions, is in fact read by a *Sophisticated* receiver as a signal of the sender's type.)

The equilibria of the reduced game are determined by the population frequencies of *Mortal* and *Sophisticated* senders and receivers.

There are two leading cases, with different implications:

- When *Sophisticated* Allies and Germans are common—not that plausible—the reduced game has a mixed-strategy equilibrium whose outcome is virtually equivalent to D-Day's without communication
- When *Sophisticated* Allies and Germans are rare, the game has an essentially unique pure equilibrium, in which *Sophisticated* Allies can predict *Sophisticated* Germans' decisions, and vice versa

In that equilibrium, *Sophisticated* Allies send the message that fools the most common kind of *Mortal* German (depending on how many believe messages and how many, like Cao Cao, invert them) and attack Normandy; while *Sophisticated* Germans defend Calais (because they know that *Mortal* Allies, who predominate in this case, will attack Calais).

(For more subtle reasons, there is no pure-strategy equilibrium in which *Sophisticated* Allies feint at Normandy and attack Calais.)

In the pure-strategy equilibrium, the Allies' message and action are part of a single, integrated strategy; and the probability of attacking Normandy is much higher than if no communication was possible.

The Allies choose their message nonrandomly, the deception succeeds most of the time, but it allows the Allies to win in the less beneficial of the possible ways.

Thus for plausible parameter values, without postulating an unexplained difference in the sophistication of Allies and Germans, the model explains why even *Sophisticated* Germans might allow themselves to be "fooled" by a costless message from an enemy.

In a weaker sense (resting on a preference for pure-strategy equilibria and high-probability predictions), the model also explains why *Sophisticated* Allies don't feint at Normandy and attack Calais, even though this would be more profitable if it succeeded.

# Application: Crawford's (2007) analysis of preplay communication of intentions in coordination games

**(preliminary paper and more detailed slides at http://dss.ucsd.edu/~vcrawfor/#Talk)**

The analysis builds on two classic analyses of explicit coordination, Farrell (1987 *Rand J*) and Rabin (1994 *JET*); henceforth "FR".

FR's models consist of a preplay communication phase followed by play of an underlying game.

Farrell studies symmetry-breaking with conflicting preferences about how to coordinate as in Battle of the Sexes (or in pure coordination games); Rabin studies coordination in a more general class of games.

FR's analyses address two conjectures:

● That preplay communication will yield an effective agreement to play an equilibrium in the underlying game

● That the agreed-upon equilibrium will be Pareto-efficient within the underlying game's set of equilibria

Regarding the structure of the communication phase, FR assume:

● Communication consists of one or more two-sided, simultaneous exchanges of messages about players' intended decisions (Rabin discusses the rationale for studying simultaneous two-sided messages rather than one-sided or sequential messages)

● The messages are in a pre-existing common language, hence understood

● The messages are nonbinding and costless ("cheap talk")

Regarding players' behavior, FR assume:

● Equilibrium, sometimes weakened to rationalizability

● Plausible behavioral restrictions defining which combinations of messages create agreements, and whether and how agreements can be changed

Under these assumptions, FR show that:

● Rationalizable preplay communication need not assure equilibrium

● Although communication enhances coordination, even equilibrium with "abundant" (Rabin's term for "unbounded") communication does not assure that the outcome will be Pareto-efficient

Equilibrium and rationalizability are natural places to start in analyses like FR's.

But the existence of an empirically successful alternative to treating deviations from equilibrium as errors makes equilibrium (or QRE) seem too strong. Rationalizability, on the other hand, may be too weak.

It therefore seems useful to reexamine FR's analyses from the point of view of a structural non-equilibrium model such as level-*k*.

Level-*k* models have not yet been tested in this setting, but their strong experimental support elsewhere makes them a natural candidate.

A level-*k* analysis allows a unified treatment of players' messages and actions and how messages create agreements; and it allows a reevaluation of FR's restrictions on how players use language.

I focus on Farrell's analysis of Battle of the Sexes, with some attention to Rabin's more general analysis. I begin with tacit coordination and then consider two-sided one-round and abundant communication.

## Tacit coordination in Battle of the Sexes

As suggested by CHC's (2004 *QJE*, Section III.C) analysis of market-entry games, a level-*k* analysis already has surprising implications for tacit coordination in Battle of the Sexes.

Subjects in market-entry experiments (e.g. Rapoport and Seale (2002)) regularly achieve better ex post coordination (number of entrants closer to market capacity) than in the symmetric mixed-strategy equilibrium, the natural benchmark.

This led Kahneman (1988, quoted in CHC) to remark, "…to a psychologist, it looks like magic". (Actually, though, it would only look like magic to a game theorist.)

CHC show that the magic can be explained by a level-*k* model: The heterogeneity of strategic thinking allows more sophisticated players to mentally simulate less sophisticated players' entry decisions and (approximately) accommodate them. They behave like Stackelberg followers, breaking the symmetry with coordination benefits for all.

The basic idea can be illustrated in Battle of the Sexes:

| | | Column H | Column D |
|---|---|---|---|
| Row | H | 0, 0 | 1, *a* |
| | D | *a*, 1 | 0, 0 |

**Battle of the Sexes (*a* > 1)**

The unique symmetric equilibrium is in mixed strategies, with $p \equiv \Pr\{H\} = a/(1+a)$ for both players.

The equilibrium expected coordination rate is $2p(1-p) = 2a/(1+a)^2$; and players' payoffs are $a/(1+a) < 1$.

In the level-*k* model, each player is one of four types, *L1*, *L2*, *L3*, or *L4*.

*L0* chooses its action uniformly randomly, with Pr{H} = Pr{D} = ½. *L1*s mentally simulate *L0*s' random decisions and best respond, thus choosing H; *L2*s choose D, *L3*s choose H, and *L4*s choose D.

| | | Column: H | Column: D |
|---|---|---|---|
| Row | H | 0, 0 | 1, *a* |
| | D | *a*, 1 | 0, 0 |

**Battle of the Sexes (*a* > 1)**

The model's predicted outcome distribution is determined by the outcomes of the possible type pairings and the type frequencies.

| Types | *L1* | *L2* | *L3* | *L4* |
|---|---|---|---|---|
| ***L1*** | **H, H** | **H, D** | **H, H** | **H, D** |
| ***L2*** | **D, H** | **D, D** | **D, H** | **D, D** |
| ***L3*** | **H, H** | **H, D** | **H, H** | **H, D** |
| ***L4*** | **D, H** | **D, D** | **D, H** | **D, D** |

Assume that the frequency of *L0* is 0, and the type frequencies are independent of player roles and payoffs (as they "should" be).

Players' level-*k* ex ante (before knowing own type) expected payoffs are equal, proportional to the expected coordination rate.

| **Types** | ***L1*** | ***L2*** | ***L3*** | ***L4*** |
|---|---|---|---|---|
| ***L1*** | **H, H** | **H, D** | **H, H** | **H, D** |
| ***L2*** | **D, H** | **D, D** | **D, H** | **D, D** |
| ***L3*** | **H, H** | **H, D** | **H, H** | **H, D** |
| ***L4*** | **D, H** | **D, D** | **D, H** | **D, D** |

Combining *L1* and *L3* and denoting their total probability $v$, the level-*k* coordination rate is $2v(1–v)$, maximized at ½ when $v = ½$.

By contrast, the mixed-strategy equilibrium coordination rate $2a/(1+a)^2$ is maximized at ½ when $a = 1$, but converges to 0 like $1/a$ as $a \to \infty$.

Thus, for $v$ near ½, empirically plausible, the level-*k* coordination rate, near ½, is higher than the mixed-strategy equilibrium rate even for moderate values of $a$, and dramatically higher for higher values of *a.*

Even though players' decisions are simultaneous and there is no actual communication, the predictable heterogeneity of strategic thinking allows some players (say *L2*s) to mentally simulate others' (*L1*s) entry decisions and accommodate them, breaking the symmetry as required for coordination, with coordination benefits for all.

The more sophisticated players become like noisy Stackelberg followers (noisy because others' types are unobservable).

The level-*k* model improves upon the mixed-strategy equilibrium by relaxing the incentive constraints requiring players to be in equilibrium.

Because *Lk* types best respond to non-equilibrium beliefs, it is natural to compare the level-*k* outcome to the best symmetric rationalizable outcome, in which players play the non-equilibrium mixed strategy $v \equiv \Pr\{H\} = ½$. When $v = ½$ the level-*k* model uses the heterogeneity of strategic thinking to "purify" this best symmetric rationalizable outcome.

Not that level-*k* thinking always makes this ideal outcome attainable: The type frequencies are behavioral parameters, not choice variables.

The level-*k* model yields a view of coordination radically different from the traditional view:

Although players are rational in the decision-theoretic sense, equilibrium—let alone selection principles such as risk- or payoff-dominance—plays no direct role in their strategic thinking.

Coordination, when it occurs, is an almost accidental (though statistically predictable) by-product of non-equilibrium thinking.

## Farrell's equilibrium analysis of Battle of the Sexes with communication, and Rabin's generalizations

In Farrell's model of Battle of the Sexes with communication, the underlying game is preceded by one or more communication rounds in which players send simultaneous messages regarding their pure-strategy intentions.

The messages are in a pre-existing common language and they are nonbinding and costless.

I denote the possible messages "h" meaning "I intend to play H" and "d" meaning "I intend to play D".

Farrell studies the symmetric mixed-strategy equilibrium in the entire game, including the communication phase, in which players take the first pair of messages that identify a pure-strategy equilibrium in the underlying game as an agreement to play that equilibrium, ignoring all previous messages.

In Farrell's equilibrium, players randomize their messages in each round until some round yields an equilibrium pair of messages, in which case they play that equilibrium; or the communication phase ends, in which case they revert to the symmetric mixed-strategy equilibrium in Battle of the Sexes.

Farrell calculates the equilibrium coordination rate with one or more rounds of communication and studies how it depends on the number of rounds.

I will describe his equilibrium by players' common values of $q \equiv \Pr\{h\}$ in each communication round and $p \equiv \Pr\{H\}$ if the communication phase ends and they play Battle of the Sexes without an agreement.

Without communication, the equilibrium failure rate is $[p^2 + (1–p)^2]$, which equals $(1+a^2)/(1+a)^2$ when $p$ takes its equilibrium value of $a/(1+a)$.

With one round of communication, the equilibrium failure rate is $[q^2 + (1–q)^2][p^2 + (1–p)^2]$, less than the rate without communication. The equilibrium $q = a^2/(1+a^2)$ so the equilibrium rate is $(1+a^4)/[(1+a^2)(1+a)^2]$.

With abundant communication, the equilibrium failure rate is a product like $[q^2 + (1–q)^2][p^2 + (1–p)^2]$, but with a separate $q$ for each round.

If the $q$s were bounded between 0 and 1, the rate would approach 0 as the number of rounds grew; but the equilibrium $q$s converge to 1 so quickly that the failure rate converges to a limit above 0 even with abundant communication.

Farrell shows that the limiting failure rate is $(a–1)/(a+1)$, and the corresponding coordination rate is $1–[(a–1)/(a+1)] = 2/(1+a)$, greater than the equilibrium coordination rate with one round. But even with abundant communication, the coordination rate $\rightarrow 0$ as $a \rightarrow \infty$.

Rabin (1994) evaluates the generality of Farrell's analysis:

- A much wider class of underlying games
- No symmetry restriction
- Richer characterization of how players use language, allowing interim agreements
- Considering the implications of rationalizability as well as equilibrium

Rabin defines notions called *negotiated equilibrium* and *negotiated rationalizability* that combine the standard notions of equilibrium and rationalizability with his restrictions on how players use language.

With abundant communication, each player's negotiated equilibrium expected payoff is at least his worst efficient equilibrium payoff in the underlying game.

Replacing negotiated equilibrium by negotiated rationalizability, each player expects (perhaps wrongly) at least the payoff of his worst efficient equilibrium.

Thus, Rabin concludes (p. 373), Farrell's insights are quite general:

"…the potential efficiency gains from communication illustrated by [Farrell (1987)] do not rely on ad hoc assumptions of symmetry or on selecting a particular type of mixed-strategy equilibrium. Rather, the efficiency gains…inhere in the basic assumptions about how players use language."

Costa-Gomes (2002 *JET*) extends Rabin's theory and tests it with the experimental data of Cooper, DeJong, Forsythe, and Ross (1989 *Rand*) and the data from Roth and collaborators' experiments on unstructured bargaining.

## A Level-*k* analysis with one round of communication

The key difficulty in analyzing two-sided level-*k* communication is extending normal-form level-*k* types to extensive-form types that determine both messages and actions.

I do this, following Ellingsen and Östling (2007), by adapting the *L0* sender type in Crawford's (2003) model of one-sided communication.

(Crawford's (2003) type hierarchy is built on a "credible" sender type, which tells the truth (there called *W0* but here called *L0*; Crawford's "credulous" receiver type *S0* is a best response to *W0*, like an *L1*).)

With two-sided communication, as Ellingsen and Östling note, a player's beliefs and best responses as a credible sender and a credulous receiver are inconsistent for sent and received messages that do not specify an equilibrium action pair, so the analysis must reconcile them in some way.

Like Ellingsen and Östling, I do this by giving priority to the credible sender type and dispensing (with regard to *L0*) with the credulous receiver type.

Thus I assume that *L0* uniformly randomizes its action, without regard to its partner's message, and sends a truthful message.

This truthful *L0* generalizes the uniform random *L0* used for games without communication and is intuitively plausible—bearing in mind that it is only the starting point for players' strategic thinking—with some experimental support from papers like those cited above. It also generalizes Crawford's (2003) truthful *W0* sender type.

In deriving types' strategies in Battle of the Sexes with two-sided communication, I assume that a type always chooses an action with the highest expected payoff, given its beliefs.

As in previous applications, I assume that payoff ties are broken randomly, so that a type chooses equally desirable actions with equal probabilities.

I also assume that the types have a slight preference for truthfulness, so that if telling the truth and lying have exactly equal payoffs, a type tells the truth.

If, in addition, both messages have equal probabilities of being true, I assume that a type sends them with equal probabilities.

With regard to types' beliefs, I assume that, because each type has a unitary model of others (*L2* believing others are *L1* etc.), it does not draw inferences about others' types from their messages.

(In Crawford's (2003) analysis, the *Sophisticated* type but not the *Mortal* (level-*k*) types draw inferences from others' messages about their types; Ellingsen and Östling assume that level-*k* types draw such inferences in their analysis of the "Poisson cognitive hierarchy" model, where types above *L1* have positive weights on all lower types.)

I also assume that if a type receives a message that contradicts its beliefs regarding its partner's action, it disregards the message and maintains its beliefs about the action, on the grounds that action preferences are stronger.

Under these and other simple assumptions, it is not hard to derive the messages for all types and the resulting coordination outcomes on the non-equilibrium path for all type pairings.

| **Type** (message) | ***L1*** (random) | ***L2*** (h) | ***L3*** (d) | ***L4*** (h) |
|---|---|---|---|---|
| ***L1*** (random) | ½H+½D, ½H+½D | D, H | H, D | D, H |
| ***L2*** (h) | H, D | H, H | H, D | H, H |
| ***L3*** (d) | D, H | D, H | D, D | D, H |
| ***L4*** (h) | H, D | H, H | H, D | H, H |

**Table 2. Level-*k* Messages and Outcomes with One Round of Communication**

"½H+½D, ½H+½D" refers to players' independently random choices in *L1* versus *L1*, which make all four possible outcomes equally likely.

For example, given *L0*'s strategy of uniformly randomizing its action and sending a truthful message, *L1* expects its partner's message to be truthful and its own message to be ignored.

*L1* therefore accommodates by choosing action D if it receives message h from its partner, and action H if it receives message d.

When *L1* chooses its own message it has not yet received its partner's message, and so it cannot predict its own action; and because *L1* expects its partner's message to be h and d with equal probabilities, both of its own messages have equal probabilities of being true.

*L1* therefore sends h and d with equal probabilities, independent of its action.

Given *L1*'s strategy, *L2* expects its partner's message to be uninformative and its own message to be believed and accommodated.

*L2* therefore chooses action H and sends message h, in each case without regard to its own or its partner's message (but if for some reason it had chosen action D instead, it would have sent message d).

## Coordination outcomes

Repeat the table for the model without communication for comparison:

| **Types** | ***L1*** | ***L2*** | ***L3*** | ***L4*** |
|---|---|---|---|---|
| ***L1*** | H, H | H, D | H, H | H, D |
| ***L2*** | D, H | D, D | D, H | D, D |
| ***L3*** | H, H | H, D | H, H | H, D |
| ***L4*** | D, H | D, D | D, H | D, D |

**Table 1. Level-*k* Outcomes without Communication**

| **Type** (message) | ***L1*** (random) | ***L2*** (h) | ***L3*** (d) | ***L4*** (h) |
|---|---|---|---|---|
| ***L1*** (random) | ½H+½D, ½H+½D | D, H | H, D | D, H |
| ***L2*** (h) | H, D | H, H | H, D | H, H |
| ***L3*** (d) | D, H | D, H | D, D | D, H |
| ***L4*** (h) | H, D | H, H | H, D | H, H |

**Table 2. Level-*k* Messages and Outcomes with One Round of Communication**

There are three notable differences between Table 1 and Table 2.

First, with one round of communication types other than *L1* always (without regard to the message sent or received) choose the action opposite to the one they choose without communication.

For example, *L2* expects its messages to be believed and accommodated, and so sends h and chooses H; but without communication *L2* expected *L1* to choose H, and so accommodates by choosing D.

Returning to the Stackelberg analogy used for tacit coordination, without communication *L1* is effectively committed (in *L2*'s mind) to choosing H; but with communication *L1* is *not* committed not to listen (because its *L0* is truthful), and this allows *L2* to use its message to take over the leadership role.

Second, in the pairing *L1* versus *L1*, there are now equal probabilities of all four {H, D} combinations, instead of the H, H outcome without communication.

This is because *L1* expects its partner's message to be truthful and its own message to be ignored.

*L1* therefore believes and accommodates its partner's message but (unable to predict which message will be true) chooses its own message randomly, so that both *L1*s end up playing H and D with equal probabilities.

*L1*'s communication skills here leave something to be desired, but its listening skills still yield a large improvement over the *L1* versus *L1* outcome without communication.

Third, in the pairing *L1* versus *L3*, *L1* still chooses H but *L3* now accommodates by choosing D.

This is because *L3* expects its partner to choose H, and so chooses D and sends d, while *L1* sends a random message but expects its partner's message to be truthful, and so ends up choosing H.

Although *L1* is not good at talking, it doesn't matter because *L3* is not listening.

The improvement here is entirely due to *L1*'s listening skills, which suffice for coordination with *L3.*

How much does one round of level-*k* communication improve coordination over level-*k* outcomes without communication or equilibrium outcomes with one round?

Focus again on the coordination rate (ignoring changes from H, D to D, H, or vice versa).

Comparing the level-*k* outcomes without communication (Table 1) and with one round (Table 2), the rate goes up from 0 to ½ for the pairing *L1* versus *L1*, from 0 to 1 for the pairings *L1* versus *L3*, and is otherwise unchanged.

If the frequencies of *L1*, *L2*, *L3*, and *L4* are $r \approx 0.4$, $s \approx 0.3$, $t \approx 0.2$, and $u \approx 0.1$ then the overall coordination rate without communication is $2(r+t)(s+u) \approx 0.48$, while with communication the overall rate goes up by $\frac{1}{2}r^2 + 2rt$, to 0.68.

Comparing the level-$k$ and equilibrium coordination rates with one round of communication, the equilibrium rate is $2(a+a^2+a^3)/[(1+a^2)(1+a)^2]$, which equals 3/4 when $a = 1$, 28/45 when $a = 2$, and converges to 0 like $1/a$ as $a \to \infty$.

Thus when $a \approx 1$ the coordination rate with one round of communication is likely to be somewhat higher for equilibrium than for a level-$k$ model (0.75 versus 0.68).

But even for moderate values of $a$, the level-$k$ coordination rate is likely to be higher than the equilibrium coordination rate.

A level-*k* analysis yields very different conclusions about the effectiveness of communication than Farrell's equilibrium analysis.

With or without communication, level-*k* coordination rates in Battle of the Sexes are largely independent of the difference in players' preferences.

By contrast, in Farrell's equilibrium analysis coordination rates are highly sensitive to the difference in players' preferences.

Unless the difference in preferences is very small, coordination rates are likely to be higher with level-*k* thinking than in Farrell's equilibria.

With one round, the analysis also justifies FR's assumption that a message pair that identifies an equilibrium leads to that equilibrium.

## A Level-k analysis of Battle of the Sexes with abundant communication

Farrell's equilibrium analysis of abundant communication assumes that players continue exchanging messages until an agreement is reached.

I assume in the spirit of Rabin's analysis that players can agree to continue for an additional round of communication by mutual consent, and that they will do so if it is mutually beneficial.

I also assume players have a slight preference for avoiding additional rounds.

Finally, I assume that players draw no inferences about their partners' types from the history of their interactions.

Under these and other simple assumptions, it is not hard to derive the messages for all types and the resulting coordination outcomes on the non-equilibrium path for all type pairings. Table 3 gives the outcomes:

| **Type** | ***L1*** | ***L2*** | ***L3*** | ***L4*** |
|---|---|---|---|---|
| ***L1*** | ½H+½D, ½H+½D if *a* < 2; 1/3H,H + 1/3D,H + 1/3H,D if *a* > 2 | D, H | H, D | D, H |
| ***L2*** | H, D | H, H | H, D | H, H |
| ***L3*** | D, H | D, H | D, D (?) | D, H |
| ***L4*** | H, D | H, H | H, D | H, H |

**Table 3. Level-*k* Outcomes with Abundant Communication**

"½H+½D, ½H+½D" refers to the uniform distribution over the four possible outcomes for *L1* versus *L1* following message pair h,h when *a* < 2.

Continuing communication can never be better for both players if their current messages already identify one of the Pareto-efficient pure-strategy equilibria in Battle of the Sexes.

By continuing they incur the slight cost of an additional round of communication, and no deviation could make that worthwhile for both of them.

This implies (finding Table 2's inefficient outcomes) that there are three kinds of type and realized message pair that might continue:

| **Type** (message) | ***L1*** (random) | ***L2*** (h) | ***L3*** (d) | ***L4*** (h) |
|---|---|---|---|---|
| ***L1*** (random) | ½H+½D, ½H+½D | D, H | H, D | D, H |
| ***L2*** (h) | H, D | H, H | H, D | H, H |
| ***L3*** (d) | D, H | D, H | D, D | D, H |
| ***L4*** (h) | H, D | H, H | H, D | H, H |

**Table 2. Level-*k* Messages and Outcomes with One Round of Communication**

● *L1* versus *L1* following one of the message pairs, d,d or h,h, that don't identify an equilibrium

● *L3* versus *L3* following its normal message pair d,d

● *L2* or *L4* versus *L2* or *L4* following their normal message pair h,h

*L1* versus *L1* following message pair d,d both expect to play H against their partner's D if communication is cut off, because each expects its partner's message to be truthful and its own to be ignored.

Given this, each is too sure of its optimistic beliefs to continue communicating.

Instead, as Rabin's analysis of negotiated rationalizability suggests is possible out of equilibrium, *L1* versus *L1* following message pair d,d both cut off communication, and so play H, H in the underlying game.

*L1* versus *L1* following message pair h,h both expect to play D against their partner's H if communication is cut off. These beliefs are too pessimistic so there is potential for improvement; but it may seem pointless to continue because they will have just failed to reach agreement in a round like the one that would ensue.

But both of *L1*'s messages have equal expected payoffs and are equally likely to be true; so if *L1*'s randomness is an unstudied response to those indifferences, the random outcomes need not be perfectly correlated each round.

Given this, the outcome if *L1* versus *L1* following message pair h,h continue will be a new random pair of messages, with a new, positive probability of identifying an efficient equilibrium (compare Costa-Gomes's (2002) "mutual grain of agreement" assumption).

If *L1* versus *L1* continue, and if the types draw no inferences from history, the process is a Markov chain with all states but h,h absorbing; the eventual outcome is either H, H; D, H; or H, D, each with probability 1/3, with expected payoff $(1+a)/3$.

If *L1* versus *L1* cut off communication, they expect to play D against H, with payoff 1.

Thus it's better to continue if and only if $(1+a)/3>1$, equivalently if $a > 2$.

The definition of *L1* gracefully overcomes what might appear an insurmountable problem in extending Farrell's equilibrium analysis of the effectiveness of abundant communication to a level-*k* model.

These models concern repeated interaction in fixed pairs, and Farrell's analysis of abundant communication inherently relies on randomness.

We are socialized to think that equilibrium players can and do consciously randomize; but it is conventional and plausible to assume that level-*k* players cannot, or at least do not, consciously randomize.

Fortunately, level-*k* players can *un*consciously randomize, and the definition of *L1* creates just the indifferences needed to make this work for *L1* versus *L1* following message pair h,h.

Summing up for *L1* versus *L1*, in the first round each of the four possible message pairs is equally likely.

If players send one of the pairs, d,h or h,d, that identify an equilibrium, then they cut off communication and play that equilibrium.

If they send d,d, then they cut off communication and play H, H.

When $a < 2$, if they send h,h, they cut off communication and play D, D.

When $a > 2$, if they send h,h they continue communicating for (at least) one more round; in that case, the eventual outcome is either H, H; D, H; or H, D, each with probability 1/3.

Like *L1* versus *L1* following message pair d,d, *L2* or *L4* versus *L2* or *L4* are too optimistic to continue communicating. They too cut off communication after the first round and play H, H in the game.

Finally, like *L1* versus *L1* following message pair h,h, *L3* versus *L3* are too pessimistic. But unlike *L1*'s messages *L3*'s are deterministic, so *L3* versus *L3* may think it's pointless to continue communicating anyway.

If they do continue, they are possibly doomed to repeat d,d forever and never reach an efficient agreement.

The only ray of hope is that, if *L3* versus *L3* do continue and there is some exogenous randomness in how messages are sent or received, or some random variation in how they learn from experience, they might eventually reach an efficient agreement by accident (such randomness is superfluous for *L1* versus *L1* following h,h; and it won't stop *L1* versus *L1* from following d,d or *L2* or *L4* versus *L2* or *L4* from cutting off communication after the first round).

## Coordination outcomes

| **Type** (message) | ***L1*** (random) | ***L2*** (h) | ***L3*** (d) | ***L4*** (h) |
|---|---|---|---|---|
| ***L1*** (random) | ½H+½D, ½H+½D | D, H | H, D | D, H |
| ***L2*** (h) | H, D | H, H | H, D | H, H |
| ***L3*** (d) | D, H | D, H | D, D | D, H |
| ***L4*** (h) | H, D | H, H | H, D | H, H |

**Table 2. Level-*k* Messages and Outcomes with One Round of Communication**

| **Type** | ***L1*** | ***L2*** | ***L3*** | ***L4*** |
|---|---|---|---|---|
| ***L1*** | ½H+½D, ½H+½D if $a < 2$; 1/3H,H + 1/3D,H + 1/3H,D if $a > 2$ | D, H | H, D | D, H |
| ***L2*** | H, D | H, H | H, D | H, H |
| ***L3*** | D, H | D, H | D, D (?) | D, H |
| ***L4*** | H, D | H, H | H, D | H, H |

**Table 3. Level-*k* Outcomes with Abundant Communication**

The outcomes with abundant communication are the same as with one round of communication, except that if $a > 2$, *L1* versus *L1* now have a coordination rate of 2/3 instead of ½; and some exogenous randomness might allow *L3* versus *L3* to raise its coordination rate above its rate of 0 with one round (the "?" in Table 3).

Updating the calibration for one round of communication, with frequencies of *L1*, *L2*, *L3*, and *L4* $r \approx 0.4$, $s \approx 0.3$, $t \approx 0.2$, and $u \approx 0.1$, if $a > 2$ the first change adds another $r^2/6 \approx 0.03$ to the overall level-*k* coordination rate with abundant communication, raising it to approximately 0.71 from the rate of 0.68 with one round and of 0.48 without communication (if $a < 2$ the rate stays at 0.68).

The second change could conceivably add as much as $t^2(1–0) = 0.06$ more, raising the coordination rate to approximately 0.77 or, if $a < 2$, 0.74.

With abundant communication, when $a > 1.94$ and possibly for lower values, the level-*k* coordination rate is greater than the equilibrium rate, $2/(1+a)$, which equals 1 when $a = 1$, 2/3 when $a = 2$, and $\rightarrow 0$ like $1/a$ as $a \rightarrow \infty$.

To the extent that level-*k* types do better than in Farrell's equilibrium analysis, they do so because, as in the level-*k* analysis of tacit coordination, the level-*k* model relaxes the equilibrium incentive constraints.

As in Farrell's analysis, the benefits of abundant communication are limited and, unless players' preferences are fairly close, most of the gains from communication are realized with only one round.
(Here, oddly, the benefits of abundant communication are more limited when *a* is small, because *L1* versus *L1* following message pair h,h then cut off communication.)

The level-*k* model's predictions are consistent with Rabin's bounds based on negotiated rationalizability, but their precision yields additional insight.

A level-*k* analysis also allows a reevaluation of FR's plausible but ad hoc restrictions on how players use language.

With abundant communication, as Rabin's analysis of negotiated rationalizability suggests, level-*k* players need not keep communicating until an agreement is reached as they do in Farrell's equilibrium.

But because "agreements" do not fully reflect the meeting of the minds that FR sought to model (instead they reflect either one player's perceived credibility as a sender or the other's perceived credulity as a receiver), a level-*k* analysis may not fully support the assumptions about agreements in Rabin's analysis of negotiated rationalizability.