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

Vincent P. Crawford

North American Economic Science Association Meetings

Tucson, 18-21 September 2007

Introduction

Recent experiments suggest that in strategic settings without clear precedents people often deviate systematically from equilibrium

The experimental evidence suggests that in such settings a structural non-equilibrium model based on level-*k* thinking (or as Camerer, Ho, and Chong (2004 *QJE*; "CHC") call it, a "cognitive hierarchy" 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 a brief introduction to level-*k* models and the experimental evidence in support of them

It then illustrates their application by using level-*k* models to resolve several puzzles regarding people's (usually experimental subjects') initial responses to novel strategic situations

The applications illustrate the generality of the level-*k* approach and the kinds of adaptations needed to use it in different settings

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 something like it) will emerge in the limit when people have had enough experience from repeated play in stable settings to learn to predict each other's responses

But in novel strategic situations, or in stable settings with multiple equilibria, we also need a reliable model of initial responses

I will argue that level-*k* models often explain more of the variation in initial responses than equilibrium or QRE, and that they are a tractable and useful modeling tool

Level-*k* models

Level-*k* models were introduced to describe experimental data by Stahl and Wilson (1994 *JEBO*, 1995 *GEB*) and Nagel (1995 *AER*), and were further studied experimentally by Ho, Camerer, and Weigelt (1998 *AER*), Costa-Gomes et al. (2001 *ECMA*), Costa-Gomes and Weizsäcker (2005), 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)

Type *Lk* anchors its beliefs in a nonstrategic *L0* type and adjusts them via thought-experiments with iterated best responses: *L1* best responds to *L0*, *L2* to *L1*, and so on

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

Lk's only departure from equilibrium is replacing its assumed perfect model of others with simplified models that avoid the complexity of equilibrium analysis; compare Selten (*EER* '98):

"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."

(Alternative specifications of level-*k* models have been considered:

● Stahl and Wilson have some higher types ("*Worldly*") best respond to 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, and for which the estimated type distribution is a useful diagnostic)

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

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

The estimated frequency of the anchoring *L0* type is usually small, so *L0* exists mainly as *L1*'s model of others, *L2*'s model of *L1*'s, 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 explanation of behavior across different settings

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 is to focus on how players model others' decisions by studying strategic thinking "uncontaminated" by learning ("Eureka!" learning is 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

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)

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

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

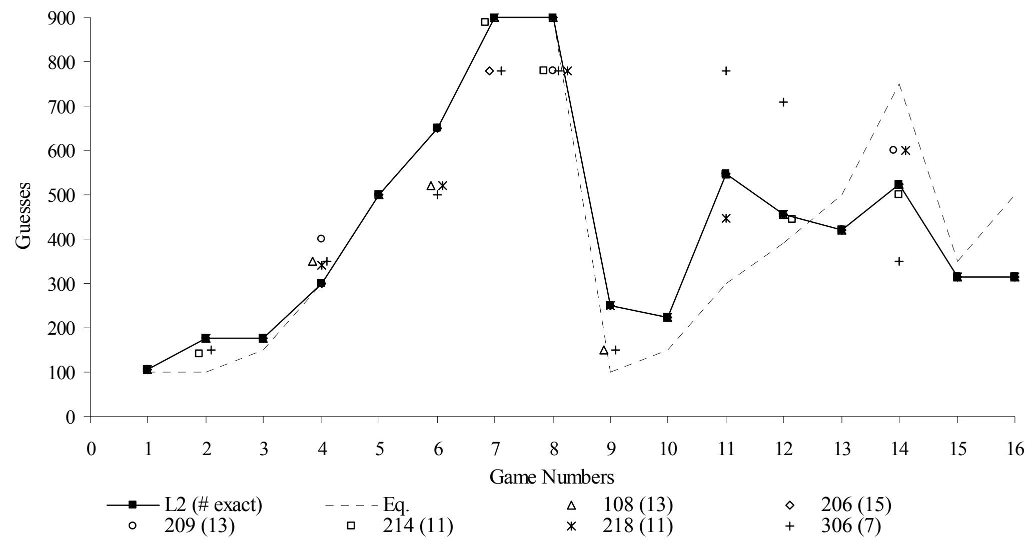
	L1	*L2*	*L3*	*D1*	*D2*	*Eq.*	*Soph.*
1	600	525	630	600	611.25	750	630
2	520	650	650	617.5	650	650	650
3	780	900	900	838.5	900	900	900
4	350	546	318.5	451.5	423.15	300	420
5	450	315	472.5	337.5	341.25	500	375
6	350	105	122.5	122.5	122.5	100	122
7	210	315	220.5	227.5	227.5	350	262
8	350	420	367.5	420	420	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

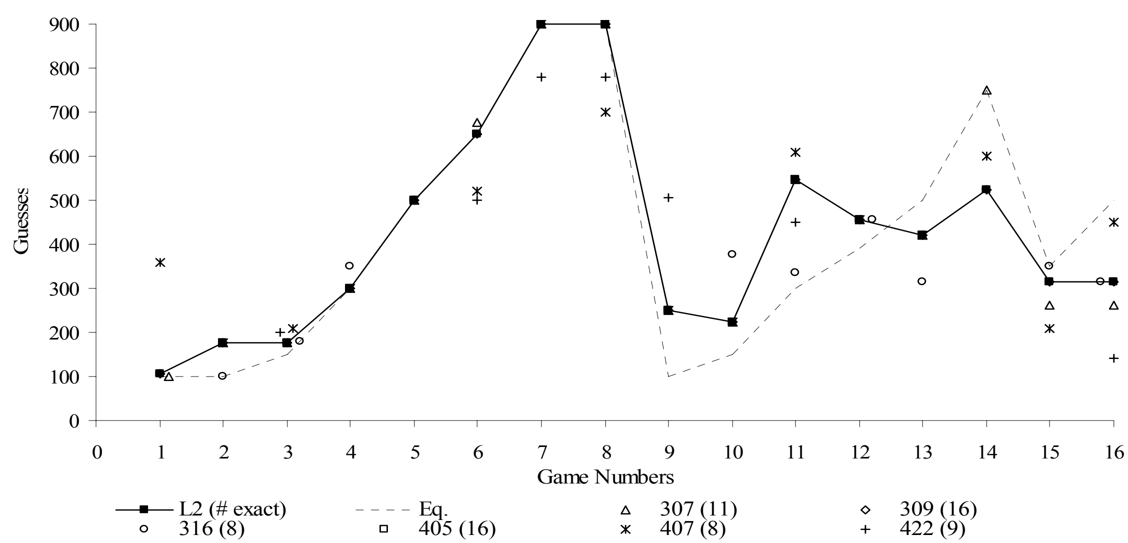
Of the 88 subjects in CGC's main treatments, 43 made guesses that complied *exactly* (within 0.5) with one type's guesses in from 7 to 16 of the games (20 *L1*, 12 *L2*, 3 *L3*, and 8 *Equilibrium*)

(The other 45 subjects made guesses that conformed less closely to a type, but econometric estimates of their types are also concentrated on *L1*, *L2*, *L3*, and *Equilibrium*, in roughly the same proportions (Table 1))

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 deviations are shown:





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

The size of CGC's strategy spaces, with from 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 1-4, we "know" that he's *L2*

Further, because the 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

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 (this is a very common interpretation of the spikes in Nagel's (1995) data)

But CGC's design separates *Lk* types from the analogous iterated dominance types (*Dk*-1, not separated from *Lk* in Nagel's design)

More detailed analysis shows that CGC's subjects are following *Lk* types that mimic iterated dominance, not doing iterated dominance

More detailed analysis also shows that CGC's subjects whose guesses are closest to equilibrium are actually following types that mimic equilibrium in some of games, not following equilibrium logic

Applications

Level-*k* models have now been used to resolve a variety of puzzles:

● CHC: coordination via structure, market-entry games, speculation and zero-sum betting, money illusion in coordination

● Blume et al. (2001 *GEB*), Kawagoe and Tazikawa (2005), Cai and Wang (2006 *GEB*), Wang et al. (2006), Sánchez-Pagés and Vorsatz (2007 GEB, 2007): “overcommunication” in sender-receiver games

● Ellingsen and Östling (2006): Aumann’s (1990) critique and why one-sided communication works better in games like Battle of the Sexes but two-sided communication works better in Stag Hunt

● Crawford, Gneezy, and Rottenstreich (2007, not yet available): coordination via focal points based on structure and framing

● Crawford (2007): coordination via preplay communication

This talk gives four illustrations, selected for their economic interest and to illustrate the modeling issues that arise in level-*k* analyses:

● CHC's analysis of "magical" ex post coordination in market-entry games (simple normal-form games with binary choices)

● Crawford and Iriberri's (2007 *AER*) explanation of systematic deviations from the unique mixed-strategy equilibrium in zero-sum two-person hide-and-seek 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 preplay communication of intentions in zero-sum games (extensive-form games)

“Magical” coordination in market-entry games

Puzzle: 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”

Resolution: 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

The more sophisticated players behave like Stackelberg followers, breaking the symmetry with coordination benefits for all

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

		Column	
		H	**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 follows one of, say, four types, *L1*, *L2*, *L3*, or *L4*, with each role filled from the same distribution

Assume (as in most previous analyses) that *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

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

Table 1. Level-*k* Outcomes

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 type) expected payoffs are equal, proportional to the expected coordination rate

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

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

For v near ½, empirically plausible, the level-k coordination rate is higher than the equilibrium rate even for moderate values of a, dramatically higher for higher values of a

Even though 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, as Stackelberg followers would (but less accurately, because others' types are unobserved)

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 equilibrium 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

Systematic deviations from equilibrium in hide-and-seek games with non-neutrally framed locations

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

Typical seeker's instructions (hider's instructions 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?*

A	B	A	A

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 analogous to the uniqueness of "*B*" but as we will see, psychologically 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)

RTH’s design 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 in a large sample from around the world, framing has a strong and systematic effect, with *Central A* most prevalent for hiders (37%) and even more prevalent for seekers (46%)

(The other boxes are chosen roughly equally often in both roles)

Folk game theory also deviates from equilibrium logic: Any game theorist would respond to the Yushchenko quote: “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?
- Hiders are as smart as seekers, on average, so hiders tempted to hide in *central A* should realize that seekers will be just as tempted to look there. 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 is asymmetric, this asymmetry of choice distributions is not explained by QRE, which coincides with equilibrium in RTH's games.)

Resolution:

A level-*k* model with a role-independent *L0* that probabilistically favors salient locations yields a simple explanation:

● 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

● 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 greater prevalence for seekers

The role asymmetry in behavior, which (despite the games' payoff asymmetry) is a mystery from the viewpoint of equilibrium, QRE, or any other theory I am aware of, follows naturally from hiders' and seekers' asymmetric responses to *L0*'s role-symmetric choices

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

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, players simultaneously and independently choose one from 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 0	0 1	0 1	1 0
2 (h)	0 1	1 0	0 1	1 0
3 (h)	0 1	0 1	1 0	1 0
J (s)	1 0	1 0	1 0	0 1

O’Neill’s Card-Matching Game

O’Neill’s game is like a hide-and-seek game, except that a 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%

Fortunately, for initial responses it turns out that there is no Ace effect

Instead there is a strong Joker effect, a full order of magnitude larger:

- 8% A, 24% 2, 12% 3, 56% J for rows
- 16% A, 12% 2, 8% 3, 64% J for columns

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

Systematic overbidding in experimental 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*))

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:

Our level-*k* 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"

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

- *Random L0* bids uniformly on the interval between the lowest and highest possible values (even if over 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 starting point of a subject's strategic thinking; 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*; 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 (in common-value auctions only)

● The bidding trade-off between the higher price paid if the bidder wins and the probability of winning (in first-price auctions only)

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 (e.g. a mixture of Random *L1* overbids and Random *L2* underbids) fits the data better than the leading alternatives

In three of the four leading cases, 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 Random *L1*, Truthful *L1*, Random *L2*, and sometimes *Equilibrium* or Truthful *L2*

(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)

The level-*k* analysis accomplishes several things:

- Provides a more unified explanation for systematic patterns of non-equilibrium bidding in private and common-value auctions

- Explores how to extend level-*k* models to an important class of incomplete-information games

- Explores the robustness of equilibrium auction theory to failures of the equilibrium assumption

- Links experiments on auctions and on strategic thinking

Preplay communication of intentions in zero-sum 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 -1	-2 2
	Attack Normandy	-1 1	1 -1

D-Day

● Attacking an undefended Calais (closer to England) is better for the Allies than attacking an undefended Normandy, and so better for the Allies on average

● Defending an unattacked Normandy is worse for the Germans than defending an unattacked Calais, and so worse for the Germans 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

Puzzle: 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 highly 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*)

		Kongming	
		Main Road	**Huarong**
Cao Cao	**Main Road**	3 -1	0 1
	Huarong	1 0	2 -2

Huarongdao

● 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 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 it 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 it

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.’ ”

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

(Cao Cao’s rationale resembles *L1* thinking, in that it assumes that the sender assumes that his message will be taken at face value

But Kongming’s rationale resembles *L2* thinking)

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?

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 boundedly rational, or *Mortal*, types and a strategically rational, or *Sophisticated*, type (interesting but rare)

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

Mortal types, like other boundedly rational types, use step-by-step procedures that generically determine unique, pure strategies, avoid simultaneous determination of the kind used to define equilibrium

Mortal types' behaviors regarding the message are anchored on analogs of *L0* based on truthfulness or credulity, as in the informal literature on deception

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 first that *Mortals*' strategies are determined independently of each other's and *Sophisticated* players' strategies, and so can be treated as exogenous (but they affect others' payoffs)

Then plug in the distributions of *Mortal* Allies' and Germans' independently determined behavior to obtain a "reduced game" between possibly *Sophisticated* Allies and 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' action, and vice versa

In this equilibrium, *Sophisticated* Allies send the message that fools the most common kind of *Mortal* German (depending on how many believe messages or, 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