Level-n bounded rationality in two-player two-stage games

Dale O. Stahl\textsuperscript{a,\ast}, Ernan Haruvy\textsuperscript{b}

\textsuperscript{a} Department of Economics, University of Texas at Austin, United States
\textsuperscript{b} School of Management, University of Texas at Dallas, United States

Received 14 December 2004; accepted 22 July 2005

Abstract

We extend the Level-n theory of bounded rationality from the domain of symmetric normal-form games to the domain simple two-player, two-stage extensive-form games. We designed and conducted experiments to test pertinent hypotheses. The extended Level-n model fits the data remarkably well and significantly better than subgame perfect equilibrium theory. Moreover, we find that the vast majority of behavior appears individualistic. Further, we characterize the non-individualistic behavior as stemming from a combination of utilitarian and Fehr–Schmidt other-regarding preferences.

© 2006 Published by Elsevier B.V.

Keywords: Level-n theory; Two-stage game; Pertinent hypotheses

1. Introduction

Experimental studies have cast doubt on the usefulness of game theoretic predictions such as subgame perfection.\textsuperscript{1} Ultimatum game bargaining indicates that people are willing to punish unfair or unkind behavior at a monetary cost to themselves. In trust, gift-exchange, and contribution games, people exhibit a willingness to incur a cost to attain a fair outcome and reward kind behavior. Possible explanations include other-regarding preferences, hot and cold emotional decisions, path-dependent phenomena, bounded rationality, and experimental procedures. Recent works (e.g., Johnson et al., 2002; Binmore et al., 2002) posit and find evidence for an alternative

\textsuperscript{\ast} Corresponding author. Tel.: +1 512 475 8541; fax: +1 512 471 3510.
E-mail address: stahl@eco.utexas.edu (D.O. Stahl).

\textsuperscript{1} For example, Güth et al. (1982), Andreoni (1988), Thaler (1988), Berg et al. (1995), Charness (2004), and Fehr et al. (1998).
explanation to other-regarding behavior in sequential games. Specifically, it appears that many
subjects are unsuccessful at backward induction. Though this is a useful insight, it is important
to identify what models of beliefs subjects apply instead.

Since the classification of behavior into hierarchies of bounded rationality has been shown
useful in numerous games and settings, and since the underlying econometric structure as a
mixture model of behavioral types easily accommodates alternative hypotheses, we extend the
Level-n theory of bounded rationality from its original domain of symmetric normal-form games
(Stahl and Wilson, 1994, 1995; hereafter Stahl–Wilson) to the domain of two-player, two-stage
games. We designed and conducted experiments to gather data to fit this model and test pertinent
hypotheses. This extended Level-n model fits the data remarkably well and significantly better
than subgame perfect equilibrium theory.

We embrace the scientific methodology of constructing alternative theories/models followed by
empirical testing. Drawing conclusions from falsifications entails the thorny problem of untangling
joint hypotheses. For example, the poor performance of subgame perfection is a falsification of the
joint hypothesis of (i) the theory of subgame perfection and (ii) the premise that the preferences
of the experiment participants were aligned with the game payoffs as specified in the theory and
implemented in the experiment. The mere appearance of other-regarding behavior implies that
the latter premise was very likely false, thereby leaving us unable to falsify subgame perfection
itself.

In our experimental design, we endeavor to implement the premise about preferences and
game payoffs, thereby strengthening our conclusions about any hypothesis conditioned on that
premise. Several features of our experimental design are intended to relieve the participants
of responsibility for others, thereby making it more likely they will seek to maximize individual
payoffs. In particular, we use a multi-task design in which no participant is obviously advantaged
or disadvantaged by the role s/he is assigned. Since competition on a level playing field is presumed
to be fair a priori, individualistic behavior is socially acceptable. Secondly, we employ binary
lottery payoffs that, by introducing an exogenous random element to final payoffs, tend to insulate
participants from responsibility for the final monetary payoffs. Third, we present the games in tree
form, therefore avoiding verbal framing via descriptive words such as “division”, “contribution”,
and so on. There have been other attempts in the literature to reduce other-regarding behavior (e.g.,
Bolton, 1991), but we believe the present approach corresponds to many real-world situations.

The tasks in our design are familiar two-stage games such as entry, ultimatum, gift-exchange,
public good contribution, and trust. Consistent with comparable studies, we find that the vast
majority of behavior appears individualistic. Nonetheless, a substantial amount of first-mover
behavior is inconsistent with subgame perfection.

Section 2 presents the extension of the Level-n model, and Section 3 presents the experimental
design and data. Section 4 gives the results of fitting the Level-n model to this data and finds

et al. (2001), Haruvy et al. (2001), Ho et al. (1998), and Gneezy (2006).
3 Diffusion of responsibility has been studied by Darley and Latane (1968), Fleishman (1980), and Barron and Yechiam
(2002). According to diffusion of responsibility, when there are many individuals in a position to help, each individual
feels less compelled to do so.
4 In the tournament approach of Bolton (1991), participants in alternating offer bargaining games were paid by their
payoff rank relative to other participants in the same role rather than directly by the amount they earned. Bolton found
this method to move behavior towards equilibrium.
5 The comparable studies, to be described in Section 3.2, use a multi-task design and discretized games.

JEBO 1935 1–19
that while the fit is reasonably good, there does appear to be other-regarding behavior. Section 5 shows how hypotheses about other-regarding behavior can be easily added to the Level-\(n\) model and tested. In particular, we find that a combination of utilitarianism and Fehr and Schmidt (1999, hereafter Fehr–Schmidt) type preferences can explain the 16% of the data that is not well-fit by the basic Level-\(n\) model.

2. Level-\(n\) bounded rationality for two-player two-stage games

The Level-\(n\) model of bounded rationality has a straightforward extension to two-player two-stage games. It postulates a population of types, called Level-\(n\) types, in which Level-0 types are uniformly random, Level-1 types believe that all other players are Level-0 and Level-2 types believe that all other players are Level-0 and Level-1 types. The proportions of each type in the population are free variables to be estimated. The simplest hypothesis is that these proportions are the same whether a player is a first-mover or a second-mover. Of course, this hypothesis can and will be tested. We start with the null hypothesis of role-independent proportions.

2.1. Second movers

In accordance with backward induction, we begin our analysis with the second movers. Since the second movers are the last players in these games, they do not need a mental model of the other players. A Level-\(n\) (\(n \geq 1\)) type simply chooses a logit best response given the available choices at each node. Letting \(\nu\) denote the precision of this logit best response, the probability of a Level-1 (for brevity) type choosing \(j\) at node \(m\) in game \(g\) is

\[
P_{gm}(j, \nu) = \frac{\exp(\nu y_{2gmj})}{\sum_k \exp(\nu y_{2gmk})}
\]  

(1)

where \(k\) ranges over the choices at node \(m\) in game \(g\), and \(y_{2gmk}\) is the respective second mover payoff (as given in Fig. 1). To accommodate unsystematic behavior and trembles, we also define a Level-0 type as being equally likely to choose each available action at each node. Letting \(\alpha_0\) denote the proportion of the second mover population that is Level-0, leaving \((1 - \alpha_0)\) as the proportion of Level-1 types, and letting \(x_{gm}(j)\) denote the aggregate second-mover choices of \(j\) at node \(m\) in game \(g\), the log-likelihood of the second-mover data is

\[
\text{LL}(x; \alpha_0, \nu) = \sum_g \sum_n \sum_j x_{gm}(j) \ln \left( \frac{\alpha_0}{2} + (1 - \alpha_0)P_{gm}(j, \nu) \right).
\]

(2)

Note also that there is no difference between Nash behavior and Level-1 behavior, so Nash types and Stahl–Wilson Worldly types are indistinguishable from Level-1 types.

2.2. First movers

Level-0 first-movers choose each first-mover branch with equal probability. Level-1 types choose a logit best response to the belief that each second mover choice is equally likely.\(^6\) Letting

\(^6\) Alternative specifications of Level-1 behavior will be discussed later.
Fig. 1. Experiment games.
ν denote the precision of this logit best response, the probability of a Level-1 first mover choosing the branch leading to node m in game g is

\[ P_1^g(m, \nu) \equiv \frac{\exp(\nu \bar{y}_1^{gm})}{\sum_k \exp(\nu \bar{y}_1^{gk})}, \] (3)

where k ranges over the branches of game g, and \( \bar{y}_1^{gk} \) is the average first-mover payoff following branch k.

Level-2 first movers believe that second movers are Level-1 types with precision parameter \( \mu \), and therefore they believe that second-mover choice probabilities are given by Eq. (1): \( P_{gm}(j, \mu) \). Hence, the first-mover’s expected payoff for choosing the branch leading to node m is

\[ E\bar{y}_1^{gm} = \sum_j [P_{gm}(j, \mu)\bar{y}_1^{gm}(j)]. \] (4)

Then, the probability that a Level-2 first mover chooses the branch leading to node m is given by

\[ P_2^g(m, \nu, \mu) \equiv \frac{\exp(\nu E\bar{y}_1^{gm})}{\sum_k \exp(\nu E\bar{y}_1^{gk})}. \] (5)

Finally, we specify a (subgame perfect) Nash type who holds subgame perfect Nash equilibrium beliefs. Subgame perfection implies that behavior in any subgame should be the Nash equilibrium

---

As noted in Stahl–Wilson (1995) allowing \( \mu \) to be a free variable is observationally equivalent to specifying Level-2 beliefs as a convex combination of \( P_{gm}(j, 0) \) and \( P_{gm}(j, \nu) \) [i.e. Level-0 and Level-1 types].
of that subgame. Accordingly, first movers who hold subgame perfect Nash beliefs anticipate that any action they choose will be followed by rational behavior. Note that in the games studied here this Nash type is equivalent to a Level-2 type with an infinite precision parameter $\mu$, as described by Eqs. (4) and (5). Hence, let $P_{g}^{\text{NE}}(m, v) \equiv P_{2}^{g}(m, v, \infty)$ denote the probabilistic choice function for a (subgame perfect) Nash type.

We do not include the Worldly type (Stahl–Wilson, 1995) who holds the belief that second movers are a convex combination of Level-0, Level-1, and Nash types, because such a Worldly type is not identifiably distinct from the Level-2 type in these two-stage games. Hence, the Level-2 type in the current model captures both pure Level-2 behavior as well as Worldly type behavior.

Let $\alpha_0$ denote the proportion of the first mover population that is Level-0, $\alpha_1$ the proportion that is Level-1, $\alpha_2$ the proportion that is Level-2 (and possibly Worldly), leaving $\alpha_{\text{NE}} = (1 - \alpha_0 - \alpha_1 - \alpha_2)$ as the proportion of subgame perfect Nash equilibrium types. Let $x_g(m)$ denote the aggregate first-mover choices of the branch leading to node $m$ in game $g$, and let $B_g$ denote the number of first-mover branches in game $g$. Then, the log-likelihood of the first-mover data is

$$\text{LL}(x; \alpha_0, \alpha_1, v, \mu) = \sum_g \sum_m x_g(m) \ln \left( \frac{\alpha_0}{B_g} + \alpha_1 P_{1}^{g}(m, v) + \alpha_2 P_{2}^{g}(m, v, \mu) + \alpha_{\text{NE}} P_{g}^{\text{NE}}(m, v) \right).$$

3. The experiment and data

Game theory assumes that a player’s preferences over potential material consequences to all players are represented by the player’s own expected utility function of those consequences and the player’s beliefs (subjective probabilities). Any regard for the consequences impacting other players is implicitly included in this expected utility function. The definition of a game gives these expected utility “payoffs” as a primitive. Under the standard interpretation, a player’s own expected utility payoffs and beliefs are sufficient to determine Bayesian rational choices. In other words, every player is by definition “individualistic” in the expected utility payoffs.

Unfortunately, this strict interpretation presents an obstacle to empirical testing of game theory, since in practice the expected utility functions are not observable. While the study of specifications of expected utility functions that incorporate other-regarding preferences, path effects, and so on is a vital research area, the testing of theory predictions conditional on the utility payoffs is also important. The latter requires players to be individualistic in the experimenter designated payoffs.

Experimental designs attempt to control for unobservables in various ways. For example, unobservable risk attitudes can sometimes be controlled for by using binary lotteries as payoffs. Group sizes of six or more have been shown to induce individualistic behavior in many-person experiments.
dictator games (Stahl and Haruvy, in press) and in public good contribution games (Isaac and Walker, 1988), which suggests that mean-matching designs with typically sized laboratory groups tend to induce individualistic behavior. In this paper we report results using a multi-task design that we suggest relieves the participants of responsibility for others, and we present the game in tree format, thereby avoiding framing effect through descriptive words such as “division”, “invest”, “accept or reject”, and so on. Several variations in the experimental protocols were used, and the behavior was found to be robust.

3.1. The experimental design

Fig. 1 shows the six two-stage games used in these experiments. The payoffs are in terms of binary lotteries, giving the probability of winning US$ 5.00 for each game. The games chosen are discretized versions of the sometimes more complex parallels in the literature. For example, offers in the traditional ultimatum game are usually continuous. The traditional gift-exchange game similarly has more choices for wage and effort levels, whereas we have two choices for each. The trust game traditionally has a continuum of choices, whereas we had two. These simplifications are necessary if these games are to be presented together in a unified fashion and in a way the subjects can and will comprehend following simple instructions. Nevertheless, we believe that the essence of each game was captured in these simple choices.

Four sessions were run at the Harvard Business School; the first three of these involved 36 participants each and the first five games listed in Fig. 1. The fourth Harvard session involved 30 participants, and the sixth game listed in Fig. 1 was substituted for the fourth game; in other words, a fifth branch with payoffs of (100, 0) was added to the ultimatum game. The Harvard subjects consisted of mostly undergraduate students from Harvard and surrounding schools (MIT, Boston University, Tufts, etc.). They were recruited through signs in these schools and through emails to subjects in the Computer Laboratory for Experimental Research database. Subjects who had participated in similar experiments were excluded. The fifth through eighth sessions were conducted at the University of Texas at Austin; those participants were upper division undergraduate and graduate students recruited campus-wide, except that Economics graduate students were not admitted since their training and experience would be atypical in the wider population.

All sessions entailed reading of the instructions and employed the strategy method, and all but the last two included an independent choice prior to the five extensive-form games. The independent task was a six-person dictator game, in which the participants were anonymously assigned to groups of six and one member of each group was anonymously picked to be the dictator for that group. Due to regulations at Harvard, individually signed receipts had to be obtained, so we were concerned that the participants would worry that their decisions could be exposed using the signed receipts. To make such linkage impossible, all sessions at Harvard were preceded by a simple lottery that gave each participant a prize of US$ 20 with probability 1/10. Neither the lottery outcome nor the dictator game outcomes were revealed prior to playing the extensive-form games, it was common knowledge that there was no linkage between the dictator game and the extensive-form games; and absolute privacy was guaranteed. Instructions are in

10 At the University of Texas at Austin, three non-participant witnesses signed an affidavit verifying the total amount of money dispersed to the session participants; thus, names, ID numbers, and so on could not be associated with any participant’s earnings or choices.
Table 1
Protocols of experimental sessions

<table>
<thead>
<tr>
<th>Session</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>Harvard</td>
<td>Harvard</td>
<td>Harvard</td>
<td>Harvard</td>
<td>Austin</td>
<td>Austin</td>
<td>Austin</td>
<td>Austin</td>
</tr>
<tr>
<td>No. part.</td>
<td>36</td>
<td>36</td>
<td>36</td>
<td>30</td>
<td>18</td>
<td>18</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>Games</td>
<td>1–5</td>
<td>1–5</td>
<td>1–5</td>
<td>1–3,5,6</td>
<td>1–5</td>
<td>1–5</td>
<td>1–5</td>
<td>1–5</td>
</tr>
<tr>
<td>Teams</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Matching</td>
<td>Mean</td>
<td>Mean</td>
<td>R. Pair</td>
<td>R. Pair</td>
<td>Mean</td>
<td>R. Pair</td>
<td>R. Pair</td>
<td>R. Pair</td>
</tr>
<tr>
<td>Dictator</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>

an Appendix C in Supplementary Material (available on the JEBO website). All lotteries were carried out independently at the end of the experiment after all decisions had been made.

In social psychology it is well known that group labels can create a “we–they” dichotomy (Messick and Mackie, 1989). Since the sequential games we investigate are natural for a two-population design (first movers and second movers), we initially called one population the Red team and the other population the Blue team. Theoretically, this Red–Blue team framing should induce a competitive environment between first and second movers neutralizing other-regarding behavior across teams, and indeed we found a preponderance of individualistic behavior. The first session also used mean matching. Each first mover was matched with each second mover, and both received the average payoff from all such matches.

To test whether the Red–Blue team framing was driving the individualistic behavior, the second session dropped the team framing, but kept everything else unchanged. This made no discernible difference in the behavior (measured by standard Chi-square tests at the 5% significance level). Next to test whether the mean matching was driving the individualistic behavior, the third session also dropped the team framing and used pairwise random matching with reshuffled pairings for each game. Again, this made no discernible difference in behavior. To test whether the multiple prizes were driving the behavior, the fourth session was like the third except that only one of the five games was selected for payment, and the binary lottery prize was increased to US$ 25; this game was selected by the random draw of a card at the beginning of the experiment, but not revealed until the end of the experiment. Once again, this made no discernible difference in behavior. Sessions five and six (at the University of Texas at Austin) replicated the second and third Harvard sessions, with no discernible difference in behavior.

Finally, to test whether the preliminary choice task in the six-person dictator game was driving the individualistic behavior, we conducted two additional sessions at UT-Austin without the preceding dictator game, but otherwise identical to session 6. Table 1 summarizes the protocols of all eight sessions. Once again, this made no discernible difference in behavior.

In total, we have 111 observations of first-mover and second-mover choices for Games 1–3 and 5, 96 observations for Game 4, and 15 observations for Game 6. Chi-square tests of the aggregate choices across the eight sessions could not reject the hypothesis that they were generated by the same process. Therefore, our subsequent analysis focuses on the pooled aggregate data.

11 The p-value for first and second movers is 0.733 and 0.368, respectively.
3.2. The data

Fig. 1 also displays the aggregate choices for each branch of the game trees. The data are remarkable in their lack of other-regarding behavior. Overall, 84.6% of all choices are consistent with subgame perfect equilibrium (indicated by thickened branches in Fig. 1): 91.9% for second movers and 65.2% for first movers. It does not follow that first movers had more regard for others since such first-mover behavior could be rationalized with individualistic preferences and non-equilibrium beliefs. Of the first-mover choices that are not subgame perfect, two-thirds are consistent with Level-1 types (indicated by “L1” in Fig. 12); in all, 88.5% of first-mover choices appear individualistic. Particularly remarkable are the ultimatum game results, which contrast with the median division of 60:40 typically observed (e.g., Roth, 1995); further, the 80:20 offer was rarely rejected.13 In the contribution, gift-exchange and trust games, there is little positive reciprocity to first-mover generosity, and in the entry and ultimatum games, there is little negative reciprocity.

It is helpful to compare the behavior in our experiments with other experiments that have used similar games and multi-task designs. Güth et al. (2001) examined a binary choice ultimatum game and found that given an equal proposed split of (10, 10) and a selfish proposed split of (17, 3), two thirds of proposers choose the equal split. Replacing the equal split (10, 10) with a slightly unequal split (9, 11) resulted in two thirds of proposers choosing the selfish outcome. This remarkable reversal indicates that without an exactly equal-split option, subjects behave selfishly. This is consistent with our result.

Falk et al. (2003) examine a binary ultimatum game with choices (8, 2) and (2, 8). Similar to Güth et al. (2001), 73% of proposers chose (8, 2). They also found that relative to a condition that allows for an equal split (5, 5) versus (8, 2), the condition without an equal split option resulted in a sharp drop of responder rejection of (8, 2) from 44.4 to 26.7%.

Note that our two ultimatum games share the no-equal-split feature with Güth et al. (2001) and Falk et al. (2003). Our ultimatum Game 4 closely resembles the Falk game, with (8, 2) and (2, 8) being the extreme points, but with two additional intermediate choices (6, 4) and (4, 6). Our somewhat more level playing field (Falk et al. had a multi-game design with four games) resulted in a higher proportion of selfish behavior on the part of proposers (79.2% versus Falk et al.’s 73%) and on the part of responders (93.6% versus Falk et al.’s 73.3%).

Results on trust games are somewhat less consistent with each other and with our results. There are many examples of binary choice trust games in various settings (e.g., Camerer and Weigelt, 1988; Cox and Deck, 2005; Engle-Warnick and Slonim, 2006; McCabe and Smith, 2000; McCabe et al., 2003). An interesting comparison, which provides some intuition for the selfish behavior we observe in our experiment, is that between McCabe and Smith (2000) and Cox and Deck (2005). McCabe and Smith studied a binary choice trust game with a first mover exit strategy of an equal split and a second mover defection strategy that gives nothing back to the first mover. In their game, second movers exhibited a low defection rate of 25%. In contrast, the same game

---

12 Except for Game 5 in which all first-mover branches are consistent with Level-1 behavior.
13 To ascertain the cause of this discrepancy, we conducted separate sessions of just the one-shot ultimatum game with (a) binary lottery payoffs with the tree presentation, (b) monetary payoffs with the tree presentation, and (c) monetary payoffs but with the typical verbal description of the game instead of the tree presentation. We found no statistically significant difference in behavior between our multi-task results and alternative designs (a) and (b). However, with alternative (c), the behavior was consistent with typical findings and statistically significantly different from the former. Therefore, it appears that the major explanation for the ultimatum game behavior in our multi-task design is due to the tree presentation.
by Cox and Deck yielded a 76% defection rate by second movers. Cox and Deck explain this stark reversal by noting that in their experiment there was a higher “social distance.” That is, the large group size in the laboratory (14–20 subjects) and the inability of the experimenter to know any individual’s actions from the amount they received resulted in a greater willingness to behave selfishly. It stands to reason that with our experimental design (with groups even larger then Cox and Deck, with a virtually double-blind protocol, and with the more level playing field), one would expect even greater rates of defection. It is also important to keep in mind that our defection action was not as extreme as in McCabe and Smith or Cox and Deck (where the first mover received 0 tokens), so our defectors may not suffer as much guilt.

In two-player entry games, the first mover (the entrant) can stay out and receive a smaller payoff than the second mover (the incumbent) or enter and potentially receive a bigger payoff at the expense of the second mover but at the risk of a costly retaliation. That is, the incumbent, upon observing entry, can retaliate by incurring some cost. Similar games, within a multi-game environment, were studied by Charness and Rabin (2002). In one such game, the first mover could stay out (400, 1200) or enter. If the first mover enters, the second mover would face (400, 200) or (0, 0). In this game, 76.9% of first movers entered and 11.5% of second movers retaliated against entrants. In our game, 75.7% of first movers entered and 6.3% of second movers retaliated. The remarkable similarity of our results to Charness and Rabin can be potentially explained by the similarity of the experimental protocol, specifically a multi-game design.

Charness and Rabin also have a “contribution” game where the first mover chooses between (700, 200) and giving the move to the second mover. The second mover chooses between (200, 700) and (600, 600); 43.8% of first movers chose to contribute, and 78.1% of second movers contributed. Remarkably like Charness and Rabin, 43.2% of our first movers contributed, but in contrast to Charness and Rabin, only 23.4% of second movers contributed. This could be due to the fact that our second movers had a choice at the end of each first mover branch, so they could punish first mover’s selfish choices.

4. Maximum likelihood estimation of Level-n model

In previous research, we found strong evidence that when fitting choice data with probabilistic choice functions of the logit form with a single precision parameter, the fit is better when the payoffs are rescaled to [0, 100] for each game. Accordingly, in the subsequent analysis, each payoff \( y \) in game \( g \) was transformed to \( 100(y - m_g)/(M_g - m_g) \), where \( M_g \) (\( m_g \)) is the maximum (minimum) payoff over both players in game \( g \). Note that this is a zero-parameter transformation that depends only on the payoffs of the game, and hence is far more parsimonious than the alternative of allowing separating logit precision parameters for each game.

Summing Eqs. (2) and (6) produces the log-likelihood function under the null hypothesis that the proportions of types and the logit precision are independent of roles. We will test this hypothesis in the next section. Maximizing this function yields \( LL = -803.99 \). As a benchmark the entropy of the data is \(-767.04\), so we have a pseudo-\( R^2 \) of 0.954. We found the MLE of \( \alpha_{NE} \) to be negligible, and upon testing we could not reject the hypothesis that \( \alpha_{NE} = 0 \); in other words, Level-2 types are indistinguishable from subgame perfect Nash equilibrium types in our data.

Table 2 gives the estimates and confidence intervals for this model.

---

14 E.g. Haruvy et al. (2001).
Table 2
Parameter estimates of Level-\(n\) model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>MLE</th>
<th>5% CI</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\nu)</td>
<td>0.173</td>
<td>0.150</td>
<td>0.204</td>
</tr>
<tr>
<td>(\mu)</td>
<td>0.400</td>
<td>0.329</td>
<td>0.529</td>
</tr>
<tr>
<td>(\alpha_0)</td>
<td>0.080</td>
<td>0.061</td>
<td>0.107</td>
</tr>
<tr>
<td>(\alpha_1)</td>
<td>0.327</td>
<td>0.276</td>
<td>0.378</td>
</tr>
<tr>
<td>(\alpha_2)</td>
<td>0.592</td>
<td>0.541</td>
<td>0.639</td>
</tr>
</tbody>
</table>

Setting \(\alpha_1 = 0\) yields a model with Level-0 and Level-2 types, the latter can be interpreted as logit subgame perfect Nash equilibrium types. The maximized LL drops dramatically to \(-885.10\); thereby strongly rejecting the restriction \(p < 0.001\).\(^{15}\) In other words, Level-1 behavior makes a significant contribution to explaining the data, so we must reject subgame perfection.

The root-mean squared error (RMSE) is 0.064 for first movers and 0.042 for second movers.

On the other hand, the Pearson Chi-square (PCS) good-of-fit statistic is 72.31, which with 32 degrees of freedom\(^{16}\) has a \(p\)-value of \(6 \times 10^{-5}\), thus, we can reject the hypothesis that the data were generated by this fitted model.

Fig. 2 displays for each game the predicted choice frequencies of each Level-\(n\) type, the prediction of the mixture model, actual choice frequencies, and three goodness-of-fit (GOF) measures: (i) the log-likelihood less the entropy, (ii) Pearson's Chi-squared (PCS), and (iii) root-mean squared error (RMSE). The Level-0 type is not shown in Fig. 2 since it is the uniform distribution over the available strategies. For second movers, since there are two branches at every such node, only the probability of choosing the left branch is displayed. Comparing the predictions of the mixture model with the actual choice frequencies, the fit is remarkably good, with four exceptions.

For second movers the fit is poor for the contribution game (2) by PCS and RMSE standards; the participants show positive reciprocity when getting more than the first mover (23% choose LA versus the model prediction of 14%), and negative reciprocity when getting less (7% choose RB versus the model prediction of 4%). Utilitarian or egalitarian preferences could help predict the positive reciprocity, but spitefulness is needed to predict the negative reciprocity.

For first movers, the fit is poor for the contribution game, ultimatum and trust games (games 1, 4 and 5, respectively): more first movers enter in Game 1 than predicted (76% versus 61%), more first movers offer close to half in the ultimatum game (18% versus 9%), and more first movers invest fully in the trust game than predicted (15% versus 8%). Utilitarian or egalitarian preferences could help predict the ultimatum game behavior, and utilitarian but not egalitarian preferences could help predict the trust game behavior. Extreme egalitarian preferences could help predict the entry game behavior.

Thus, it appears that the departures from the predictions of the fitted Level-\(n\) model cannot be explained by a single other-regarding preference type. In the next section, we present a parsimonious model of this apparently other-regarding behavior.

\(^{15}\) Restricting \(\alpha_0 = 0\) also, would be a test of the stronger hypothesis of a population made up of only subgame perfect types, and that hypothesis is even more strongly rejected.

\(^{16}\) From Fig. 1, it is obvious that there are 14 independent first mover decision nodes and 18 independent second mover decision nodes.
It is also interesting to note from Fig. 2 that Level-2 first movers in the gift-exchange game (3) are somewhat ambivalent about choices A and B rather than strongly preferring the subgame perfect choice B. The reason the Level-n model predicts this is due to the small payoff difference for the second mover following A, implying that Level-2 first movers expect a significant proportion of second movers to choose LA (as if reciprocating).

5. A Level-n model with other-regarding behavior

An attractive feature of the underlying mixture structure of the Level-n model is the ease with which alternative types can be introduced and tested. In particular, in addition to the Level-0 type, and the individualistic Level-1 and Level-2 types, suppose there is an other-regarding (OR) type and let $\alpha$ denote the latter proportion in the population.

For ease of presentation, we invoke the "projection" hypothesis of psychology: that a person tends to model others like oneself. Specifically, we assume that an individualistic type believes the other players are individualistic, and we assume that the OR type (being forward looking) believes all other types are Level-1 types with OR preferences and precision $\mu$. Appendix A in Supplementary Material provides statistical tests that confirm these assumptions.

17 See, for example, Croson and Miller (2003).
Table 3
Parameter estimates of OR Level-n model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>MLE</th>
<th>5%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \nu )</td>
<td>0.230</td>
<td>0.196</td>
<td>0.278</td>
</tr>
<tr>
<td>( \mu )</td>
<td>0.383</td>
<td>0.328</td>
<td>0.458</td>
</tr>
<tr>
<td>( \alpha_0 )</td>
<td>0.021</td>
<td>0.009</td>
<td>0.045</td>
</tr>
<tr>
<td>( \alpha_1 )</td>
<td>0.221</td>
<td>0.170</td>
<td>0.279</td>
</tr>
<tr>
<td>( \alpha_2 )</td>
<td>0.600</td>
<td>0.548</td>
<td>0.645</td>
</tr>
<tr>
<td>( \alpha_3 )</td>
<td>0.158</td>
<td>0.118</td>
<td>0.192</td>
</tr>
</tbody>
</table>

Deferring the specification of OR preferences until later, let \( Q^2_g(m, \nu, \mu) \) denote the OR first-mover logistic choice function with precision \( \nu \) given the belief that the second-mover has OR preferences and precision \( \mu \). Also, let \( Q_{gm}(j, \nu) \) denote the OR second-mover logistic choice function with precision \( \nu \) for game \( g \), node \( m \) and strategy \( j \). Weighting by preference type, for a random subject, the probabilistic choice function in the first-mover role is given by

\[
P_{FM}^g(m) \equiv \alpha_0 B_g + \alpha_1 P^1_g(m, \nu) + \alpha_2 P^2_g(m, \nu, \mu) + \alpha_3 Q^2_g(m, \nu, \mu),
\]

(7a)

and in the second-mover role by

\[
P_{SM}^{gm}(j) \equiv \alpha_0 2 + (\alpha_1 + \alpha_2) P^1_{gm}(j, \nu) + \alpha_3 Q_{gm}(j, \nu).
\]

(7b)

We now come to the challenging problem of specifying the OR preferences. From Section 4, instances of poor fit suggest that two types of OR behavior are present: utilitarianism to account for the amount of trust in Game 5, and spite to account for negative reciprocity in Game 2. To capture such behavior, we use a two-parameter model in the spirit of Fehr–Schmidt:

\[
U_1(y_1, y_2) = y_1 + \theta_1 \min\{y_2 - y_1, 0\} + \theta_2 \max\{y_2 - y_1, 0\},
\]

and

\[
U_2(y_1, y_2) = y_2 + \theta_1 \min\{y_1 - y_2, 0\} + \theta_2 \max\{y_1 - y_2, 0\},
\]

(8)

where \( \theta_1 \geq 0 \) and (for convexity) \( \theta_2 \leq \theta_1 \). Utilitarian preferences are represented by \( \theta_1 = \theta_2 = 1/2 \). Spiteful preferences are represented by \( \theta_2 < 0 < \theta_1 \). As an initial hypothesis, we assume that \( \theta_1 = 1/2 \) and that \( \theta_2 = 1/2 \) or \( -1 \) with equal probability. Thus, with one additional parameter (\( \alpha_3 \)), we have an encompassing model with individualistic and OR preferences.

The maximized LL of this encompassing model is \(-782.99\), an increase of 21.00 which has a \( p \)-value less than \( 9 \times 10^{-11} \). Therefore, this addition of OR preferences makes a very statistically significant contribution to explaining the data. The pseudo-\( R^2 \) is a remarkable 0.980. Moreover, the PCS goodness-of-fit statistic decreases to 32.72, which has a \( p \)-value of 0.431; therefore, we cannot reject the hypothesis that this model is the data generating process. Finally, the RMSE decreases to 0.032. Table 3 gives the parameter estimates and confidence intervals for this model.

18 There are several other prominent models of fairness (e.g., Bolton, 1991; Bolton and Ockenfels, 2000). We chose this specification because it nested several other models, as will be explained shortly.

19 This is equivalent to specifying two types of other-regarding preferences, each equally likely.

20 As in the previous section, we also tested and concluded that Level-2 types are indistinguishable from subgame perfect Nash equilibrium types in our data.
The estimate of the new parameter ($\alpha_3$) indicates that about 16% of the population behaves as if they have OR preferences, which is consistent with our analysis of poor fits when OR preferences are excluded. Conversely, about 84% of the population behaves as if they have individualistic preferences. In other words, our experimental design appears to have successfully induced the intended game-theoretic payoffs for 84% of the participants.

It is also noteworthy that the estimated proportion of Level-0 types decreases substantially and the estimated precision parameter ($\nu$) increases, because the way a model with only individualistic preferences tries to fit OR behavior is as random error (i.e. a larger $\alpha_0$ or a smaller $\nu$). The proportion of individualistic Level-2 types remains essentially the same at about 60%. Thus, allowing for OR behavior reduces the estimated proportions of Level-0 and individualistic Level-1 behavior, but not individualistic Level-2 behavior.

Fig. 3 displays the predicted choice frequencies of the Level-1, Level-2, and OR types for each game, as well as the prediction of the mixture model, and the three GOF measures (similar to Fig. 2). Compared to the Level-$n$ model without OR types, the overall fit is better, especially in those games for which the former model fits poorly. For first movers in Game 1 (entry), the increased precision of Level-2 types raises the predicted frequency of entries (from 61 up to 68%), and the PCS test now passes. In the ultimatum games (games 4 and 6), the spiteful FS types strongly prefer the 60–40 split, which makes the mixture prediction closer to the observed frequencies. In the trust game (5), the utilitarian types prefer the maximal investment, while the more egalitarian Fehr–Schmidt types prefer the intermediate level of investment; both behaviors make the mixture prediction closer to the observed frequencies.

For second mover in the contribution game (Game 2), both OR types exhibit positive reciprocity when getting more than the first mover, which increases the mixture prediction from 14 to 20%; at the other node (when the second mover is getting less than the first mover), the spiteful OR types are indifferent between LB and RB, whereas all other types (except Level-0) strongly prefer LB. This increase in predicted frequency of RB choices is offset somewhat by the decreased estimate of the proportion of Level-0 types, so the mixture prediction for negative reciprocity (RA) increases only slightly from 4 to 5%. Nonetheless, the PCS declines dramatically (from 12.27 to 1.77) and the RMSE declines (from 0.074 to 0.027). In all other games, the fit generally improves due to the replacement of Level-0 types with OR types. For instance, in the ultimatum games (games 4 and 6), the spiteful FS type rejects offers of 80–20 or worse and accepts all better offers; clearly substituting this behavior for the Level-0 behavior fits the data better.

We tested the restrictions on OR preferences that $\theta_1 = 1/2$, and that $\theta_2 = 1/2$ or $-1$ with equal probability by considering an encompassing model with two subtypes of OR preferences: (i) utilitarian ($\theta_1 = \theta_2 = 1/2$), and (ii) general Fehr–Schmidt ($\theta_1 \geq 0$ and $\theta_2 \leq \theta_1$). Further, we allowed the relative proportion of these subtypes to be a free parameter. The maximized LL increased by only 0.67, which with 3 degrees of freedom has a $p$-value of 0.72. Therefore, we cannot reject the restrictions entailed by our specification of OR preferences.21

To test the validity of assuming the same proportion of types for both first-movers and second-movers, we fitted the model, allowing independent proportions for each role. The increase in the maximized LL was 2.40, which with three degrees of freedom22 has a $p$-value of 0.188. Therefore, we cannot reject the hypothesis of a common set of parameters for both roles.

21 Alternative specifications of OR preferences such as CES and Cox and Friedman (2002) made no improvement; see Appendix B in Supplementary Material.

22 The first-mover model has the full five parameters, while the second-mover model has only three parameters (no $\mu$ and no $\alpha_2$); thus, the unrestricted model has eight parameters, while the pooled model has five parameters.
Fig. 3. Level-n with other-regarding types: predicted and actual choice frequencies and goodness-of-fit (GOF) measures. The three GOF measures reported are (i) the log-likelihood less the entropy, (ii) PCS, and (iii) RMSE.

The overall earning performance of each type is the expected payoff that it would have received when facing the actual empirical distribution of play in the data set. Sophisticated learning (e.g., rule learning as proposed by Stahl, 2000) and evolutionary dynamics would predict that those types with above average earning performance would increase their share of the population, while those types with below average earning performance would decrease their share.23 Table 4 gives this overall performance measure24 for the five types; since the performance calculated separately

23 We are assuming, as in G¨uth et al. (2000), that survival fitness depends on the material payoffs rather than the utility payoffs for the other-regarding types as well as the individualistic types.

24 In rescaled payoffs, which does not affect the relative performance comparisons.
Table 4
Relative performance of types

<table>
<thead>
<tr>
<th>Game</th>
<th>Level-0</th>
<th>Level-1</th>
<th>Level-2</th>
<th>Util.</th>
<th>FS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>55.18</td>
<td>55.21</td>
<td>60.16</td>
<td>55.40</td>
<td>60.55</td>
</tr>
<tr>
<td>2</td>
<td>65.26</td>
<td>65.28</td>
<td>82.06</td>
<td>63.57</td>
<td>55.55</td>
</tr>
<tr>
<td>3</td>
<td>20.35</td>
<td>24.12</td>
<td>24.01</td>
<td>16.47</td>
<td>24.37</td>
</tr>
<tr>
<td>4</td>
<td>38.64</td>
<td>63.05</td>
<td>63.61</td>
<td>47.04</td>
<td>44.00</td>
</tr>
<tr>
<td>5</td>
<td>35.42</td>
<td>40.23</td>
<td>47.16</td>
<td>28.55</td>
<td>40.50</td>
</tr>
<tr>
<td>6</td>
<td>28.80</td>
<td>33.92</td>
<td>46.52</td>
<td>33.47</td>
<td>33.31</td>
</tr>
<tr>
<td>Average</td>
<td>42.70</td>
<td>48.79</td>
<td>54.94</td>
<td>41.84</td>
<td>44.71</td>
</tr>
</tbody>
</table>

for first movers and second movers are very similar, only the combined performance measure is reported here. The column labeled “Util.” stands for the utilitarian component of the OR type, while “FS” stands for the Fehr–Schmidt component with $\theta_1 = 1/2$, and $\theta_2 = -1$. The “Average” row is computed by weighting each game by the number of subjects.

The Level-2 type (which leads to subgame perfect behavior) is the highest overall performer, and hence we would expect to see the Level-2 share increase in a subsequent period. The Utilitarian type and the Level-0 type are the poorest performers, and hence we would expect to see their shares decrease. This change in the population would make the Level-2 type perform even better in subsequent periods. Consequently, rapid convergence to the subgame perfect equilibria is quite likely.25

### 6. Discussion

We extended the Level-$n$ theory of bounded rationality from the domain of symmetric normal-form games to the domain simple two-player, two-stage extensive-form games. We designed and conducted experiments to test pertinent hypotheses. The extended Level-$n$ model was found to fit the data remarkably well and significantly better than subgame perfect equilibrium theory. We found that about 62% of the population behavior is consistent with Level-2 behavior, about 22% is consistent with Level-1 behavior, and about 2% is random Level-0 behavior. The remaining 16% of the behavior is consistent with a combination of utilitarian and Fehr–Schmidt other-regarding preferences. These proportions apply to both first and second movers, which is understandable since the experimental design made it likely that subjects would form a single belief regarding other types and the potential for future repercussions.

We presented the OR extension to satisfy the anticipated curiosity of readers and to gain some insight about the behavior that is not well-fit by the individualistic Level-$n$ model. That behavior contains elements of utilitarianism and spitefulness. We presented a model with individualistic Level-0, Level-1, and Level-2 types and one hybrid OR type that was either utilitarian or spiteful with equal probability. The improvement in all goodness-of-fit measures was substantial and significant, and we could not reject the hypothesis that the data were generated by the fitted model.

We are surprised that the simple Level-$n$ model of first movers does so well. Based on past studies of the ultimatum game, which found near equal splits to be the modal choice, we expected that the naive Level-1 belief that second movers are equally likely to make any choice at every

25 This conjecture has been confirmed in a later study.
node would be unreasonable since such belief would lead to the first mover demanding all or most of the pie. Instead, we contemplated an alternative conceptualization of second-mover strategies in terms of reservation levels: accept any offer equal to or exceeding, say, $r$. Then, if each such reservation level is equally likely, the first mover’s optimal proposal would be a 50–50 split. In contrast, the 80–20 split is the predominant first-mover proposal in our data, consistent with the naïve Level-1 type.

Also pertinent to explaining our ultimatum behavior is the large proportion of Level-2 types (about 60%) in comparison to Stahl–Wilson (1995), and Haruvy et al. (2001). However, since Level-2, naïve Nash and Worldly types are indistinguishable in our extensive-form data, the comparison should be with the sum of those types in the previous studies, and that sum is around 60%.

We did not attempt to improve the fit by allowing the parameters to vary across games because (1) the fit is already remarkably good and (2) without a theory that links specific games to specific parameter values, having a six-fold increase in free parameters would have no predictive value. In contrast, our results characterize behavioral regularities across a variety of games that have predictive value.

A large body of experimental literature has demonstrated that utility functions may be more complex than selfish monetary considerations alone. Unfortunately, it may take many more years before all possible non-pecuniary considerations are mapped and game theoretic predictions can be tested with these new ‘complete’ utility functions. Until then, we have argued that proper testing of game theoretic predictions requires experimental designs in which participants care only about their own payoffs. We found that participants were predominantly individualistic in our design. The maximum-likelihood results for both the individualistic Level-$n$ model and the extended model with OR types suggest that our design succeeded in inducing the intended game-theoretic payoffs for at least 84% of the participants. Further, behavior appears quite robust to many variations of the multi-task design, including team framing, mean-matching versus random pairwise matching, and single versus multiple prizes; Clearly, more research is needed into what design features are necessary and sufficient for predominantly individualistic behavior. Whatever the outcome of that research, however, our results on the performance of the Level-$n$ model will stand.

Uncited reference


Appendix A. Supplementary data


References


betting conjecture. Mimeo.
Stahl, D., Haruvy, E. Other-regarding preferences: egalitarian warm glow, empathy, and group size. Journal of Economic
Behavior and Organization, in press.
and Organization 25, 309–327.
Behavior 10, 218–254.