RMM Vol. 0, Perspectives in Moral Science,
ed. by M. Baurmann & B. Lahno, 2009, 199–206
http://www.rmm-journal.de/

Gary E. Bolton and Axel Ockenfels

Testing and Modeling Fairness Motives*

Abstract:
The advent of laboratory experiments in economics over the last few decades has produced an enormous literature devoted to describing, testing and modeling economic and social behavior. Measured by publications and citations, the development of social preference models to capture decisions motivated by fairness and other social criteria, is one of the success stories in this literature. But with this success, and maybe even because of it, controversies have arisen about what the models can and cannot do. In this note, we comment on some of these debates. Our main theme is that descriptive models of behavior should be judged with respect to their *usefulness*. This is often neglected, partly because there are no accepted measures and tests for the usefulness of a model, while standard procedures to test whether a model is *true* are readily available. A model that does not capture a 'grain of truth' is unlikely to be useful; however, the relationship is not monotonic in that a 'truer' model is not necessarily a more useful model.

1. Are Fairness Models True?

We organize our discussion around a simple fairness model called ERC (Bolton and Ockenfels 1998; 2000). The model characterizes how people trade-off material self interest with a preference for fair distribution. More specifically, the model stipulates that each agent has a 'motivation function' such that for a given relative payoff (defined as one's own payoff relative to the total payoff allocated in a reference group), an agent's choice is consistent with the standard assumption made about preferences for money; that is, more money is better than less. Alternatively, holding the pecuniary payoff fixed, an agent's motivation function is strictly concave in one's relative payoff, with a maximum around the allocation at which one's own share is equal to the average share. That is, people care about their status in a reference group and, in particularly, may dislike an unfavorable relative position. Otherwise, the model is consistent with standard game theoretic modeling, including the usual assumptions about rational behavior and equilibrium play. The model organizes a large and otherwise disparate set of laboratory observations, many anomalous to models based solely on self interest, such as equity seeking behavior in bargaining games and reciprocity

* Bolton gratefully acknowledges financial support from the National Science Foundation. Ockenfels gratefully acknowledges financial support from the Deutsche Forschungsgemeinschaft.

in social dilemma games. But the model also implies the very competitive play that economists expect, and that is observed, in market games (see Bolton and Ockenfels 2000 for details).

Is the model true? The model basically traces social behavior back to the decision makers' concern for relative position. However, human behavior in general and social behavior in particular is a complex phenomenon. E.g., fairness does not only have a motivational side, but also cognitive, biological (including neurobiological but also chemical, physical etc.), sociological, adaptational and other roots. It has been shown, for instance, that intranasal administration of oxytocin causes a substantial increase in trust among humans, thereby greatly increasing the benefits from social interactions (Koesfeld et al. 2005). Also, subjects who briefly held a cup of hot (versus iced) coffee judged a target person as having a 'warmer' personality (generous, caring), and subjects holding a hot (versus cold) therapeutic pad were more likely to choose a gift for a friend instead of for themselves (Williams and Bargh 2008). There is also a large and convincing literature showing that humans do not behave fully rationally, but rather follow boundedly rational heuristics such as that suggested by Simon's satisficing approach.

ERC focuses on individual motivations and their interactions with strategic decision making in laboratory games, and neglects any cognitive, biological or sociological roots of social decision making. For example, it cannot capture the observation that holding a cup of hot coffee affects social behavior. As a consequence, it cannot be 'true' in the sense of capturing all factors that may be relevant.

2. All Models Are Approximations

A 'true' model of social decision making needs to incorporate all relevant motivational, cognitive, biological, sociological and other factors. This is unfeasible and probably undesirable. We use models as maps.[1] You use a different map depending on where you want to go and how you are going to get there (walk, drive or fly). The most useful maps communicate the critical information in compact form, and so they omit or even sometimes actively distort facets of the landscape. We propose to judge and test models using the same standards we judge maps by: On the basis of how accurately they portray an aspect of the landscape we want to know about.

Observe, for instance, that a critical ingredient in a subway map's success is the simplifying assumptions it makes about the locations of the train stations: Stations on the London Underground map, for example, are laid out like nodes on a grid, either vertical, horizontal or at a 45 degree angle to one another. The result is a clear and economical tool for navigating the underground. You can quickly see that to get from Russell Square to Great Portland Street,

[1] The discussion in this section follows Bolton (forthcoming).

you should change trains at King's Cross. Yet if you took this map as a guide for foot travel, you would end up walking from Russell Square to Great Portland Street in the wrong direction. Setting the Underground and geographically correct maps side-by-side, it is easy to see the value of this distortion—for subway riders: Restricting station locations to a grid structure makes the map far more transparent and simpler to use. We think we want our models to be simple for similar reasons—and the cost for this simplicity, as it is for maps, is a loss of detail and sometimes even some distortion.

Increasingly, models are how findings from the economics lab are communicated to the larger community of researchers. This is how it should be. All of these models are approximations of what we have learned. There is, as there should be, vigorous debate about which model provides the best approximation. But just as with a map, there are inevitably trade-offs between accuracy and simplicity, and in particular, between breadth of use and detail. A map of a city university campus tends to show the locations of laboratory and classrooms in greater detail and accuracy than the non-university buildings surrounding them. A map of the entire city, however, will generally have less local detail and accuracy but broader application. So you will use the city map to get from the airport to the university subway stop but use the university map to find your way from the subway stop to the economics department.

3. Simplicity Has Value

In a recent paper, Bergh (2008) addresses what he deems a puzzle concerning the "huge impact" of ERC and related work by Fehr and Schmidt (1999, hereafter FS). Bergh aims to critically examine "the merits of the models as a theory of fairness and explanations of human behavior". His basic premise is that "simply put, a scientific explanation of a phenomenon needs to provide an answer to the question of why it occurs". He cites evidence that conflict with the models. In the concluding discussion, he returns to ask "why a theory with rather limited applicability and no deeper explanatory power has become so widely popular and heavily cited". One potential explanation he offers is that "the theory was easy to incorporate in other sub-disciplines, where it could be used to seemingly explain central questions". In fact, we would say that is exactly the point: Models like ERC are successful precisely because they provide a simple and useful map, in this case of how relative standing can influence decision making. Adding cognitive, biological or other explanatory factors to the model would probably make it 'truer' or 'deeper' (if these terms can be adequately defined), but not necessarily more useful.

To illustrate the point, we consider an example also cited by Bergh (2008), a paper by Engelmann and Strobel (ES 2004) reporting laboratory tests of fairness models. ES reject ERC and FS as an explanation for their data in favor of an explanation that involves a preference for efficiency combined with self interest and maximin (a fairness measure that makes somewhat different predictions

than the measures used by ERC or FS). This is basically the same combination of preferences proposed by Charness and Rabin (2002). Engelmann and Strobel go on to argue that a preference for efficiency, beyond what self interest can explain, may play a more prominent role in the broader set of games to which ERC and FS are typically applied. So, ES reject ERC and FS by claiming that they focus on wrong motives.[2]

However, we think ES's paper, and recent papers like it, are indicative of confusion over what it is that social preference models are trying, and should be trying, to achieve. As Alvin Roth (2002) puts it, "since we know that approximations aren't precisely true, it is easy not to be impressed by evidence that they are not". ES may have shown that ERC and FS are wrong in the sense that they do not get every possible lab experiment right. But, all models are approximations. This also holds for ES's model, and it is easy enough to identify laboratory games within exactly their own laboratory environment (which we will not repeat here in detail) in order to show that their model, too, is an approximation.

Consider, for instance, an ES kind of game in which the decision maker has to choose one of two payoff distributions, A and B, over a total of six subjects. He himself gets paid 8 regardless of his choice, and the other five subjects get 8, 8, 8, 15, 1, respectively, for alternative A, and 2, 2, 2, 33, 2, respectively, for alternative B (all payoffs in Euro). Alternative B strictly maximizes *both* maximin and efficiency, whereas alternative A is the ERC choice. In an experimental study of this situation, 45 out of 48 subjects (94%) chose alternative A.[3] Does this failure disqualify maximin plus efficiency as a possible explanation of other games? We don't think so (although, the experiment suggests that subjects avoid efficiency when it comes with costs to others). Social utility models cannot, and we would say should not, aim to capture every behavior in all settings. Rather, the challenge is to identify general principles of economic behavior that are useful in organizing and predicting decision patterns. Our little experiment rejects maximin plus efficiency as the 'true' explanation but it does not assess the usefulness of this approach. For this, one needs to analyze a wider, non-degenerate range of economically salient situations. ERC and FS did exactly this.[4]

---

[2] See Bolton and Ockenfels 2006 for a detailed reply to Engelmann and Strobel 2004.

[3] Analogous to ES we kept the decision maker's payoff fixed, played this game in strategy method, and explicitly informed subjects about the fact that distribution B yielded a higher maximin payoff and higher efficiency. We did not, however, choose a three-person game, because this would have not allowed us to keep the decision-maker's payoff fixed and to increase maximin *and* efficiency while at the same time diminishing fairness as measured by ERC or FS.

[4] Regarding salience, in the ERC paper we dealt with gaming situations that economists have been traditionally concerned with, having to do with markets, bargaining, public goods, and the like; games where the decision makers face meaningful trade-offs. The games ES examine, on the other hand, involve a single decision maker who chooses one of three payoff allocations for three people. In 8 of the 11 treatments, the decision maker had *no* payoff at stake. Both ERC and FS models admit strictly self interested behavior. This means that for nearly three quarters of the treatments in ES's study, every choice available is consistent with both ERC and FS. The salience critique also applies to the few games where decision maker stakes do vary with the decision, since the expected value of the differentials (given that there was only a one third chance a decision makers choice would count for payoff) are never greater than DM2/3 or about US$0.30. Moreover, in 2 of

ES's response to their rejection of ERC and FS is that we need to develop models that incorporate yet more motives; a logical response if you think the goal is to explain all behavior in a single model. But, as we demonstrated above, the question of which combination of motives should be included is far from obvious, even within the restricted setting of ES's experiment. To further illustrate the point, if we go beyond ES's setting, their approach becomes basically unmanageable, because the number of motives tends quickly to pile up. For instance, one important place where self interest, maximin and efficiency unambiguously fail is the ultimatum game, where all three of these motives imply that no positive offer should be rejected, contrary to a mountain of data. As ES put it in an earlier version of their comment, ultimatum game behavior "is only consistent with a model based on efficiency, maximin preferences, selfishness, competitiveness, and perceived intentions if the role of inequality aversion is relatively weak compared to intentions and competitiveness". Clearly, models based on such a large number of motives as suggested by ES are unattractively complicated, intractable in more complex situations, and risk becoming tautological and less robust. Summing up, they are not useful.

The challenge to the researcher is not to test which model is 'true', because all models are approximations and so it is easy not to be impressed that they are not. There is no hope that any tractable model can fully capture the complexities of human decision making. Even if we focus only on motivations, things are quickly getting complex; e.g., Selten (1990, 653) states: "There is no reason to suppose that human behavior is guided by a few abstract principles. Nobody should be surprised if it turns out that the motivational system is as complex as the anatomy and physiology of the human body." So, at least at this stage of knowledge, the challenge to social decision making research is to find tractable models that capture important drivers of social behavior in useful ways.

This view has implications for testing and modeling fairness. Most importantly, there is value in simplicity. A subway map that includes details about the rolling stock, signaling, communication, power supply, fare collection, air-conditioning systems, tunnel corrosion, noise, vibration, floating slabs, and the temperature of the drivers' coffee is not useful to travelers. Even if a perfectly true model of the subway could be devised, one that is identical to the real world subway, it would not be useful to most of us. Coming back to fairness in laboratory games, ERC is arguably one of the simplest formulations of the idea that

---

the 3 treatments where selfishness could matter, the selfish choice is taken, respectively, by just 10% and 23.3% of the subjects. (In the third treatment, the selfish choice is taken by a majority, but the choice agrees with *all* the fairness and efficiency measures considered.) This poor showing for selfishness is very different than what we observe in most other experiments, including in experiments where fairness and reciprocity are important. But it has an easy explanation: There is little opportunity in ES's game to express self interested behavior. ES also want to convince the reader that people care about efficiency as well as fairness. To an economist, the way you show you care is by paying. In the only two treatments where efficiency is distinctive from the choice of a purely self-interested subject, the expected 'price' for efficiency is never greater than DM1/3 or about US$0.15 (and the efficient choice is also the fair choice as measured by maximin). There is not a single decision in the ES experiment that implies a subject is willing to pay *any* positive amount to increase efficiency.

people care about status in social and economic interaction. So, if there is value in simplicity, testing models against ERC need to discount more complex models. For instance, a natural comparison for ES's maximin plus efficiency is with explanations of equal parsimony. Just as we can pair efficiency and maximin as ES did, we can pair the other explanatory variables and compare to see what fits best. If we do this, a combination of ERC plus maximin, two fairness motives, explains more choices in 6 of ES's treatments and less in none, than does maximin plus efficiency. In some treatments, the improvement in predictive power is substantial; for example, in treatment E in ES, accuracy nearly doubles, from 39.7 percent to 76.4 percent. In other words, and contrary to a central claim by ES, efficiency is not critical to explaining the data. That said, it is not our intention to push any particular explanation too hard: ERC plus efficiency also does better than maximin plus efficiency. Combinations with FS do pretty well. ERC plus FS explains some things other pairs do not.[5]

A quote from a recent survey by Cooper and Kagel (forthcoming) on other-regarding preferences dealing with ERC and FS succinctly summarizes our point: "It was clear at the time that both these papers were written that they had to be 'wrong', but as one of my old teachers used to say 'wrong in the right way'."

4. Conclusions: What Fairness Models Can Explain

The fairness models have attracted a good deal of interest and, together with the work of a number of others, have triggered a new and larger wave of research on the role of fairness and reciprocity in economic decision making. In our view, the useful insights behind ERC and FS are three: First, simple measures of fairness can approximate the fairness behavior of a population of people over a broad set of games. Heretofore, many economists thought that individual answers to the question of 'what-is-fair' were too diffuse for fairness to be a useful predictor of behavior. Of course, not everyone measures fairness the same way. The claim is that the measure offered by ERC provides a good approximation the fairness driven behavior of people in broad set of scenarios. Second, fairness can explain acts of reciprocity. Fairness—sometimes in the guise of justice, sometimes in the guise of equity—and reciprocity are age old preoccupations of people everywhere, and these models make explicit the intimate link between them. Third, certain types of institutions, such as competitive markets, induce people

[5] Nor does investigating the various motives separately support ES's views. In the only treatments in the experiment in which efficiency (study 3, treatment Ey), respectively maximin (study 3, treatment R), makes a prediction distinctive from the other non-selfish motives on the table, the distribution of choices made cannot be distinguished from random (p = .272 and .684, respectively, Chi-squared test). Directly comparing the fairness measure proposed by FS and ERC, ES claim that FS performs better (because, according to ES, FS is more in line with maximin concerns). But in saying this, ES restrict themselves to study 1, where FS performs better in 3 out of 4 games. Yet, ERC does better in all other games of the other two studies that yield distinctive predictions, and so overall explains a larger percentage of all choices than FS in 5 out of 8 cases.

to behave *as if* they are completely materially self interested. Hence, it is not that peoples' concerns for fairness are confined to, say, political or legal spheres, rather the institutional structure shapes the expression of these concerns. For some institutions, a model that takes material self interest as the sole driver of behavior sacrifices little accuracy.

At the same time, ERC and other models are not 'true' in the sense of capturing all relevant factors of social decision making. In fact, the controls in the laboratory allow to 'falsify' *any* model yielding testable predictions, just because by definition approximations are not exactly true. In our view, the most promising challenge is to better understand the role of procedures in social decision making. Simple models like ERC have directed interest towards this and related research questions. In Bolton, Brandts and Ockenfels (1998), we designed the first laboratory study in experimental economics investigating the role of 'intentionality', and in Bolton, Brandts and Ockenfels (2005) we designed the first laboratory study in experimental economics investigating the role of 'procedural fairness' in social behavior. We also believe that more research is needed regarding the role of reference group and reference point formation in social behavior (Bolton and Ockenfels 2005). Beyond such motivational aspects of decision making, it is an exciting endeavor to complementarily investigate the cognition, biology and sociology behind social behavior, and to put the findings to real world tests (Bolton, Greiner and Ockenfels 2009; Bolton and Ockenfels 2008; Ockenfels 2009). That said, we believe that a *concern for one's relative position* will persist to take a central role in our understanding of social decision making: a model that turns out to be useful is likely to capture a grain of truth.

References

Bergh, A. (2008), "A Critical Note on the Theory of Inequity Aversion", *The Journal of Socio-Economics* 37, 1789–1796.

Bolton, G. E. (forthcoming), "Testing Models and Internalizing Context: A Comment on Vernon Smith's 'Theory and Experiment: What Are the Questions?'", *Journal of Economic Behavior and Organization*.

—, J. Brandts and A. Ockenfels (1998), "Measuring Motivations for the Reciprocal Responses Observed in a Simple Dilemma Game", *Experimental Economics* 1, 207–219.

—, — and — (2005), "Fair Procedures: Evidence from Games Involving Lotteries", *Economic Journal* 115, 1054–76.

—, B. Greiner and A. Ockenfels (2009), "Engineering Trust—Reciprocity in the Production of Reputation Information", Working paper, University of Cologne.

— and A. Ockenfels (1998), "Strategy and Equity: An ERC-Analysis of the Güth-van Damme Game", *Journal of Mathematical Psychology* 42, 215–26.

— and — (2000), "ERC: A Theory of Equity, Reciprocity and Competition", *American Economic Review* 90(1), 166–193.

— and — (2005), "A Stress Test of Fairness Measures in Theories of Social Utility", *Economic Theory* 25(4), 957–82.

— and — (2006), "Measuring Efficiency and Equity Motives: A Comment on 'Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments'", *American Economic Review* 96(5), 1906–11.

— and — (2008), "Does Laboratory Mirror Behavior in Real World Markets? Fair Bargaining and Competitive Bidding on eBay", Working Paper Series in Economics, No. 36, University of Cologne.

Charness, G. and M. Rabin (2002), "Understanding Social Preferences with Simple Tests", *The Quarterly Journal of Economics* 117(3), 817–869.

Cooper, D. and J. Kagel (forthcoming), "Other Regarding Preferences: A Selective Survey of Experimental Results", in: Handbook of Experimental Economics.

Engelmann, D. and M. Strobebl (2004), "Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments", *American Economic Review* 94(4), 857–869.

Fehr, E. and K. M. Schmidt (1999), "A Theory of Fairness, Competition, and Cooperation", *Quarterly Journal of Economics* 114, 817–868.

Kosfeld, M., M. Heinrichs, P. J. Zak, U. Fischbacher and E. Fehr (2005), "Oxytocin Increases Trust in Humans", *Nature* 435, 673–676.

Ockenfels, A. (2009), "Marktdesign und Experimentelle Wirtschaftsforschung", *Perspektiven der Wirtschaftspolitik* 10, 31–53.

Roth, A. E. (2002), Slides Prepared for Al's "Experimental Economics" class, mimeo.

Selten, R. (1990), "Bounded Rationality", *Journal of Institutional and Theoretical Economics* 146, 649–58.

Williams, L. E. and J. A. Bargh (2008), "Experiencing Physical Warmth Promotes Interpersonal Warmth", *Science* 322(5901), 606–607.