A Model for Rabin Fairness under Risk

Michael Golz, Morgan Cheatham, Andreas Karagounis

1 Introduction and Literature Review

In his December 1993 paper "Incorporating Fairness into Game Theory and Economics," Matthew Rabin proposed a new model to account for reciprocity.[1] Namely, Rabin formulated a new utility function that accounts for people's perception of the kindness or unkindness being shown to them given the actions of others, and demonstrated the equilibrium behavior suggested by his new model. He begins his paper by making explicit the intuitions he sought to incorporate in a model of reciprocity. Firstly, people will sacrifice their own well-being to help those who are kind to them. Secondly, people will sacrifice their well-being to be unkind to those who are unkind to them. Finally, both these behaviors change as the cost of the sacrifice changes. However, though he lists a number of natural extensions for his model, the paper is primarily concerned with expounding the baseline model and analyzing equilibria.

We have a very natural question as an extension of Rabin's original model that we cannot find to have been formalized in the literature in a satisfying way. Our goal is to incorporate intuitive notions of people's behavior under risk. In doing so, our contribution to the literature is a simple foundation for new analysis of reciprocal behavior under risk. In 2007, Koszegi, Botond, and Rabin produced a new paper regarding reference-dependence under risk.[2] As will be explained below, aside from our dissatisfaction with the complexity of the model suggested in that paper, we are interested in modeling ideas that are much less sophisticated, and hence conducive to a cleaner setup. Our model is (to our knowledge) novel in its style of manipulation of Rabin's original work.

Rabin's 1993 model assumes complete information and deterministic outcomes to players' actions. Our objective is to relax the determinism of the model and answer the question, "Can we model how perceived kindness changes in circumstances of risk in which players are subject to probabilities of losing initial endowments, and are there implications for equilibrium behavior?"

The question can be motivated in a number of ways. For example, in a 2015 paper, W. Kip Viscusi and Ted Geyer discuss an interesting problem arising in the allocation of public resources under the EPA Superfund Program.[3] They suggest that political actors responsible for allocating public funds to remediate contaminated land sites overallocated such resources. This occurs, they suggest, because political actors are subject to conservatism bias, and hence over-exaggerate the importance of low probabilities of bad health outcomes resulting from

[1] Rabin, Matthew. "Incorporating Fairness into Game Theory and Economics." American Economic Review (1993): 1281-302. Web.

[2] Koszegi, Botond, and Matthew Rabin. "Reference-Dependent Risk Attitudes." American Economic Review 97.4 (2007): 1047-073. JSTOR. Web.

[3] Viscusi, W. Kip, and Ted Gayer. "Behavioral Public Choice: The Behavioral Paradox of Government Policy." SSRN Electronic Journal (n.d.): n. pag. Mercatus Center, Mar. 2015.

contaminated sites. Their explanation is grounded in a particular type of bias regarding risky outcomes.

We wonder if there are other explanations for such behavior. In a game theory environment, these political actors might instead be responding to an implicit understanding of what is considered fair given the reported probabilities. If they recognize that what their constituents consider to be a fair allocation given the constituents' expected utility is distorted by the risk of loss that the constituents are facing, they may simply be responding to multi-stage incentives to pump up the expected payoff they offer. If the amount of kindness perceived to have been shown is distorted by the probability of loss, what players must offer to be considered kind will change as the risk changes.

Thus, instead of a model in which actors are subject to conservatism bias, we might suggest a model in which over-allocation is the result of political actors' cognizance that perceived kindness is relative to probabilities associated with the game. Our goal is to model distortions in the actual utilities of involved players, rather than upward exaggerations of the reported probabilities alone. As a remark, while the applications to game theory are natural and the model can be evaluated using games, the paper is concerned with incorporating social behavior under risk into individual's utility functions, and is thus primarily behavioral in nature; we do not intend to model the issue formally in the context of game theory.

Implementing stochastic choice environments into the Rabin Fairness Model requires further consideration of framing and reference-dependence. While the Rabin Fairness Model incorporates variables that depend upon the choice environment, such as the perception of the level of fairness in a given choice, the model fails to provide a complete frame. Without thoroughly accounting for framing, Rabin's model cannot account for reference-dependence. Instead, the Rabin utility function is dependent upon the kindness perceived by final wealth positions, in accordance with expected utility theory (Schmidt).[4] The determinant nature of the Rabin Fairness model proposes a sufficient explanation for the lack of framing in the model, given that perceived kindness is not calculated in a stochastic environment over risky choices; however, existing theoretical frameworks can provide supplementary insight to contest the insufficient framing in Rabin's model. Specifically, prospect theory offers insight into decision-making across probabilistic alternatives involving risk. From Kahneman and Tversky's (1979) development of prospect theory, the significance of final wealth positions is replaced with value functions representing losses and gains.[5]

In its original form, however, prospect theory falls short by violating first-order stochastic dominance in the transformation of individual probabilities into decision weights (Schmidt). A useful addendum to prospect theory, namely cumulative prospect theory, ameliorates this violation by enabling the translation of cumulative probabilities into appropriate decision weights.[6] Perhaps a more relevant contribution of this model in regards to framing and reference-dependence lies in the value function, which depends not on final wealth, but gains and losses relative to

[4]Schmidt, U. (2003). Reference dependence in cumulative prospect theory. Journal of Mathematical Psychology, 47(2), 122-131. doi:10.1016/s0022-2496(02)00015-9.

[5]Kahneman, D., and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Emmitsburg, MD: National Emergency Training Center.

[6]Tversky, A., and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty J Risk Uncertainty, 5(4), 297-323. doi:10.1007/bf00122574.

a reference point or status quo. In this setting, both the reference point of an asset and the deviation from the reference point determine the output the value function.

Koszegi and Rabin expand upon cumulative probability theory and other contemporary models for reference-dependence in their 2006[7] and 2007 (mentioned above) papers detailing a model for reference-dependent preferences in stochastic environments. The model captures the essence of reference-dependence over stochastic reference points such that an individual's reference point could be determined by the expectation of an outcome from the reference lottery. According to their model, the outcome provides gain or loss to the individual based on the relationship between outcome and the reference lottery. In opposition to other frameworks for understanding reference-dependence through value functions, Koszegi and Rabin implement a utility function that is dependent on consumption utility $m(c)$ and a gain-loss utility $n(c|r)$ conditional on the reference lottery. Finally, Koszegi and Rabin's model makes two key assumptions: 1. an individual's expectations about the outcome of the reference lottery are defined by rational, probabilistic beliefs held shortly before an outcome is generated, such that a personal equilibrium is reached when the individual correctly predicts the choice environment and their response, and 2. The reference point is the status quo.

While this framework proposes a nuanced method for modeling reference-dependent risk attitudes, Koszegi and Rabin's model suffers from fundamental shortcomings. First, the model does not account for cumulative probability weighting, or "what outcomes a person pays attention to," as defined in their 2007 paper. In a reciprocal choice environment where there exists some probability p of experiencing loss for one or both players, individual distortion of p would inherently affect the expected outcome from a given lottery, and consequently, individual choices. Thus, a satisfying treatment of risky situations necessitates a way of accounting for how people actually map probabilities into decision making. If people treat a 10% change differently in different ranges of the probability space, it has consequences for behavior. We must weight probabilities in the way that the decision makers do.

Additionally, the reference dependent nature of Koszegi and Rabin's model is primarily motivated by consumption and considerations of gains versus losses. These two features serve as the reference points. However, it is plausible that these are not the only relevant factors in a person's reference space. Namely, we posit that the actual lotteries themselves play an integral role in a person's point of reference. In social interactions that involve risk, we think that there are substantive differences in how an individual who is subject to a high probability of loss perceives the actions of others as opposed to an individual subject to a low probability of loss. This effect is independent of how the change in expected payoffs influences their perception.

To demonstrate how severe this shortcoming is, consider a simple example. Suppose you are about to board a flight back to T.F. Green Airport in Providence. Upon landing, you are planing to take the airport bus back to downtown, after which you can walk back to your house. However, you are told that there is a chance that the flight will be landing 30 minutes late. You know that this is past the time which the bus stops making trips. So

[7]Koszegi, B., and Rabin, M. (2006). A Model of Reference-Dependent Preferences. The Quarterly Journal of Economics, 121(4), 1133-1165. doi:10.1093/qje/121.4.1133.

you contact a friend, tell him the chance that the plane will be late, and ask if he would make himself available to pick you up from the airport in case your flight gets back too late to catch the bus. Your friend can choose to either help you out or blow you off.

Your perception of the kindness of your friend's choice is highly dependent on the chance that the plane lands late. For example, if it is only a 5% chance and he decides not to help out, you might be miffed but in reality probably don't blame him that much for blowing you off. However, if there is a 50% chance your plane will arrive late, you might be more upset. You think that your friend's action is quite obtuse given his knowledge that a mere coin flip stands between you and getting home. Of course, the critical thinker might suggest that probabilities are accounted for simple by adjusting the expected utilities to the new lottery. A mathematical example in section four will show that the change in expected utility that results from a change in probability does not alone explain much of the behavior we seek to model.

The applications of this model are numerous. For example, in the provision of public services, one can imagine a scenario in which political actors allocate capital from a common endowment to their constituents who are subject to a risky environment. Other applications for this type of risky choice under settings of probabilistic perception of kindness are prevalent; however, the essential questions driving the development of this model were initially motivated by scenarios of provisioning public services.

2 Some Notation

$X = (x, y, z)$ where x is player i's action, y is the action player i's believes player -i will take, and z is the action i thinks j believes i will take.

$w = (endowment_1, endowment_2, endowment_{1,2})$ is the vector containing the initial endowments of each player and, if applicable, a common endowment such as public resources that might be allocated to either player through some type of game. Since we are addressing risky situations which involve a probability of loss, we must define what could be lost and, if applicable, what could recuperate such loss: i.e. we need initial conditions.

$\rho = (q, p)$ is the vector contains player 1's probability of losing his initial endowment, q, and player 2's probability of losing his initial endowment, p. Hence, $\rho_1 = q$ and $\rho_2 = p$.

$E(\pi_i(X, w, \rho))$ is the expected payoff to player i given the actions and beliefs in X, the initial endowments w, and probabilities in ρ.

$\bar{f}_{-i}(X, w, \rho)$ is player -i's kindness to player i given the actions and beliefs prescribed in X and the initial en-

dowments w.

$f_i(X, w, \rho)$ is player i's kindness toward player -i given the actions and beliefs prescribed in X and the initial endowments w.

We use the same formulation of the functions f and $\bar{f}$ as in the original Rabin model, except that the payoffs are in expectation. Namely, the kindness function takes the difference of the expected payoff and the equitable expected payoff (halfway between the expectation of the most player i could give player -i given that -i plays y and the least non-Pareto dominated outcome i could give to player -i, given that -i is playing y), and normalizes it by the range of possible expected payoffs that could be offered.[8] Note that if a player is not subject to a probability of loss, the expectation operator is trivial.

$$f_i(X, w, \rho) = \frac{E(\pi_{-i}(X, w, \rho)) - E(\pi^e_{-i}(X, w, \rho))}{E(\pi^{high}_{-i}(X, w, \rho)) - E(\pi^{min}_{-i}(X, w, \rho))}$$

$C_\alpha(p, v) : p \in [0, 1] \mapsto [0, 2]$ is a weighting function that maps probabilities to scale factors. In this paper, we use a fusion of the transformed Heavyside function and the functional form from Prelec [1998]; α prescribes the value of the parameter in the Prelec probability weighting function. As will be described below, v is the value of the original kindness function, though only the sign matters.

$\bar{C}_{\alpha i}(\rho, v) : \rho \in [0, 1]^2 \mapsto [0, 2]$ is a function that takes the difference of two of the above-mentioned weighting functions evaluated at the two player's probabilities of loss: $1 + C_\alpha(\rho_i, v) - C_\alpha(\rho_{-i}, v)$.

$u_i(X, w, \rho) : X \mapsto \mathbb{R}$ is i's utility function that maps actions designated in X to the real numbers given the initial endowments and probabilities of loss. In this paper, we work primarily with utility money which we take to be linear, though this could be changed without any difficulty.

3 The Model

Our proposed model follows from a set of intuitive notions about how people likely respond in situations involving a risk of loss. These notions are as follows:

a) The perceived kindness of a given outcome produced for one player by the actions of another player decreases as the probability of loss increases, likely in a non-uniform manner. If you offer me just barely enough money to recuperate my loss in expectation, that will be perceived as less kind if my chance of loss is 90% rather than if it is 10%.

[8] McNeill, John. "Other Regarding Preferences 2." Brown University, Providence. 7 Apr. 2016. Lecture.

b) In circumstances in which both players are subject to a probability of loss, what is considered kind is highly reference dependent. Suppose Player 2 has a probability of loss of .5. The kindness of Player 1 to Player 2 would be higher if Player 1's probability of loss is .45 than if it was .1. What a player perceives as kind changes relative to the gap between the two players' probabilities of loss.

c) There is likely asymmetry in how each player distorts the probability space depending on whether it is the other player or himself that is experiencing the loss. In other words, a given outcome is perceived to be less kind by the player on the receiving end than it is perceived to be by the player on the sending end.

We will speak often about "recuperation" by which we mean the extent to which a player recuperates his expected losses by way of the outcome produced by the actions of the players, such as by way of an offer from the common endowment by the opposing player. As described above, this language is representative of the public goods motivation for the project, but is widely applicable to many circumstances.

Moving ahead, we restrict our modeling to circumstances involving two players, and leave it to future iterations to expand to cases involving more players. We begin by considering situations in which only one player is subject to a risk of loss. Note that we borrow heavily from Rabin 1993 for notation. Refer back to the section on notation if you are unfamiliar with the Rabin setup; concepts such as the equitable payoff are used in exactly the same way, except that there is now the notion of expected payoff. We treat utility as linear in money, and consider only basic expected payoff, again noting the opportunity to expand the model to more nuanced representations of utility.

3.1 Case: Only 1 Player Subject to Risk of Loss (1PR)

Consider the following situation in which two players are interacting, where only one player is subject to some risk of loss. Both have initial endowments of capital given in the vector w. Player 2 is subject to a probability p of losing his endowment. Based on the actions of the players, Player 2 may be be recuperated for some of his losses, such as through the offering of capital from a common public endowment. Player 1's action will usually involve some payoff to himself depending on how recuperative funds are allocated to Player 2. For example, you might imagine Player 1 to be a politician who is allocating the public budget to an initiative which might recuperate some of his constituents given some risk factors they face.

We would like to model notion (a) and (c) in the utilities of the players. To do so, consider the following utility functions (again refer to the notation section for definitions of constituent parts):

$$u_1(X, w, \rho) = \pi_1(X, w, \rho) + \bar{f}_2(X, w, \rho)(1 + f_1(X, w, \rho)C_\beta(p, f_1))$$
$$u_2(X, w, \rho) = \pi_2(X, w, \rho) + (\bar{f}_1(X, w, \rho)C_\alpha(p, \bar{f}_1))(1 + f_2(X, w, \rho))$$

We note the substantive differences between these functions and Rabin's 1993 model. Firstly, the f functions include as arguments the initial endowments and the probability of loss. These allow us to calculate the expected

payoff to the players given the actions and beliefs in X. For example, suppose Player 1 offers some of the common endowment to Player 2, and keeps whatever he does not offer. If $w = (10, 10, 20)$ and $p = .25$, an offer to Player 2 of \$10 in recuperative funds from the common endowment by Player 1 results in Player 1 having $10 + 10 = \$20$ total, and Player 2 having an expected payoff of $10 + [(.25) * 0 + (.75) * 10] = \$17.5$.

The other important difference is the corrective function within the kindness terms. The function in Player 2's utility represents our attempt to incorporate notion (a). That is, we needed to penalize the kindness of a given offer from Player 1 to Player 2 depending on the probability of loss which Player 2 faces. The corrective term we use is $C_\alpha(p, v)$ where

$$C_\alpha(p, v) = 1 + (2H(-v) - 1)e^{-(-ln(p))^\alpha}$$

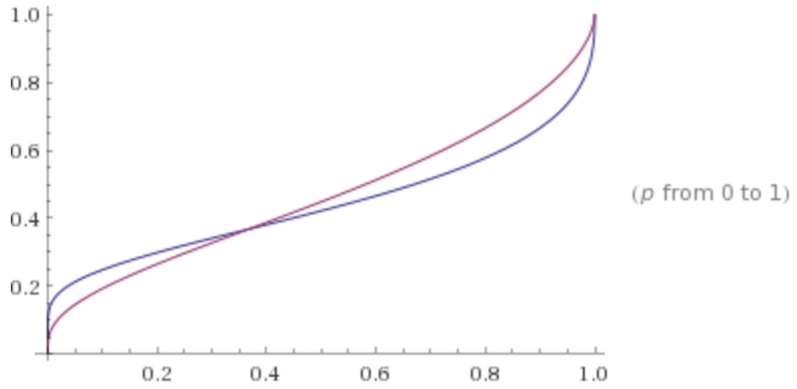
Here, H(x) is the Heaviside step function. Hence,

$$2H(-v) - 1 = \begin{cases} 1 & v < 0 \\ -1 & v > 0 \\ 0 & v = 0 \end{cases}$$

To break this down, we reiterate that we have set out to create a model that corrects downward the perceived kindness of any given outcome when a player is subject to a probability of loss. In other words, we want to penalize the kindness function, so we must be able to scale both positive and negative values downward. The Heaviside function, transformed as above, allows us to do this. If the other player is being kind to me, $v > 0$, so $0 < C_\alpha(p, v) < 1$. However, if the other player is being unkind to me, then $v < 0$, and $1 < C_\alpha(p, v) < 2$. Since the corrective function is multiplied by the value of the kindness function itself, unkindness is made to look more unkind (negative values are multiplied by a number greater than 1) and kindness looks less kind (positive values are multiplied by a number between 0 and 1). It has been demonstrated empirically that many people distort the probability space, so we utilize the probability weighting function from Prelec [1998], which has the benefit of mapping probabilities back into the probability space while emphasizing changes in probabilities very near to 0 or 1.[9] This captures the essential non-uniform nature in notion (a) of how kindness is perceived across the probability space: there is more erratic behavior near the tail probabilities, but overall monotonically increasing weights. The plot below shows Prelec's function graphed for $\alpha = .4$ (blue) and $\alpha = .6$ (red).[10]

$$f(p) = e^{-(-ln(p))^\alpha}$$

[9] Prelec, Drazen. "The Probability Weighting Function." Econometrica 66.3 (1998): 497. JSTOR. Web.
[10] Wolfram—Alpha. Wolfram Alpha LLC, n.d. Web. 09 May 2016.



In essence, what the corrective function is doing is penalizing a given level of kindness by using the Heaviside function to scale both positive and negative values downwards. For $p = 0$, we recover the original Rabin model without risk.

The other important feature is the corrective term that appears in Player 1's utility function. It operates in much the same manner as the term in Player 2's utility function, except that we assume that $\beta > \alpha$. This means that changes in the probability of loss are considered more nearly linear by Player 1 than by Player 2. A given change in the probability of loss near the tails is much more extreme from Player 2's persepctive (blue line), who is experiencing the loss, than for Player 1 (red line), who in a sense is acting in a more coolly rational manner as to what a change in probability actually represents. This asymmetry between the corrective terms is our attempt to handle notion (c).

3.2 Case: Both Players Subject to Risk of Loss (2PR)

In the case that both players are subject to some probability of loss, we have to account for a degree of reference dependence. This is the substance of notion (b). If a player knows the other is subject to some probability of loss, he will update what he considers kind or unkind. The amount which he considers a given recuperative offer unkind is lessened if he recognizes the other player is also acting under circumstances with a personal risk of loss. In formulating the two person case, we must account for the relative nature with which a player determines the fairness of the outcomes.

Now we consider a situation in which two players are interacting, and in which each is subject to some probability of loss, where this probability of loss is likely different for each player. Both have initial endowments of capital given in the vector w. Player 2 is subject to a probability p of losing his endowment; Player 1 is subject to a probability q of losing his endowment. The players engage in behavior that may have self-recuperative effects, mutually recuperative effects, or both.

The first modification to the 1PR case is that Player 1's utility now contains a corrective term for the kindness shown to him by Player 2, and Player 2 now has a corrective term regarding the kindness he shows to Player 1. The more substantive modification arises due to our desire to address the reference-dependent nature of 2PR. To

do so, we introduce a new form, $\bar{C}_\alpha(\rho)$, where

$$\bar{C}_{\alpha i}(\rho, v) = 1 + C_\alpha(\rho_i, v) - C_\alpha(\rho_{-i}, v)$$

This means that the scale factor is dependent on both players' probabilities of loss. If they are subject to the same probability, we recover the original Rabin model. If you have a lower probability of loss than I do, it masks the kindness of your actions. If you have a higher probability of loss than I do, it makes any kindness look better and "explains" any unkindness: unkind actions seem less malevolent in light of your difficult position of balancing between showing me kindness and protecting yourself from more certain loss. The extreme cases are compelling. For example, if I am guaranteed to lose, and you have no probability of loss but are being unkind to me, it doubles what I perceive as your unkindness. If we reverse roles and you are guaranteed to lose but are being unkind to me, the unkindness cancels out.

$$u_1(X, w, \rho) = \pi_1(X, w, \rho) + (\bar{f}_2(X, w, \rho)\bar{C}_{\alpha 1}((\rho), \bar{f}_2)(1 + f_1(X, w, \rho)\bar{C}_{\beta 1}((\rho), f_1))$$
$$u_2(X, w, \rho) = \pi_2(X, w, \rho) + (\bar{f}_1(X, w, \rho)\bar{C}_{\alpha 2}((\rho), \bar{f}_1)(1 + f_2(X, w, \rho)\bar{C}_{\beta 2}((\rho), f_2))$$

Having presented the functional form and intuitions of the model, we now move on to demonstrate what kinds of behaviors it would explain by showing the model's power to shift equilibria in interactions between two players.

4 Demonstration of a Change in Equilibrium

Thus far, the form of our suggested utility functions has been motivated solely by intuitions of social behavior under risk. We have drawn upon established mathematical and empirical work to formulate a modification to an existing model, but have only alluded to its potential to drive equilibrium outcomes in interactions between agents. In this section, we offer a tangible scenario that demonstrates the ability of our model to map these intuitions to true outcomes. We present a game as a toy model to show what type of behavior ought to result from the above model. We present the game for only the 1PR case, but the method is easily extended to the 2PR case.

The scenario we are considering is the following. There are two players, who have the endowment vector $w = (2.5, 2.5, 2.5)$. Player 2 is subject to a probability of loss $p = .25$ in one setup, and $p = .75$ in another. Player 1 offers an amount from the common endowment, either a high (\$2) or low (\$.75) amount. Simultaneously, Player 2 decides to accept or reject the forthcoming offer. If he accepts, he receives the offered amount, and then a random process determines if he does or does not lose his initial endowment. If he rejects, both players are left with their initial endowment, and the random process is run to determine if Player 2 does or does not lose his initial endowment.

This might seem pathological without the dynamic element of Player 2 hearing the offer before he decides,

but the game suffices to demonstrate the mechanics of the model. The experimental section will add back into the mix a dynamic nature to the scenario. For now, consider the matrices below which contain the expected payoffs for each player (for Player 1, expected payoff is trivial, as his outcomes are deterministic). We use standard expected utility (linear in money). The matrices are for $p = .25$ and $p = .75$ respectively. The best responses are boxed.

		Player 2	
		A	R
Player 1	H	$(3, \boxed{3.875})$	$(\boxed{2.5}, 1.875)$
	L	$(\boxed{4.5}, \boxed{2.625})$	$(\boxed{2.5}, 1.875)$

		Player 2	
		A	R
Player 1	H	$(3, \boxed{2.625})$	$(\boxed{2.5}, .625)$
	L	$(\boxed{4.5}, \boxed{1.375})$	$(\boxed{2.5}, .625)$

As can be seen, there is a unique pure-strategy Nash Equilibrium in which Player 1 offers low and Player 2 accepts. This equilibrium is true of both the $p = .25$ and $p = .75$ scenarios. Of particular interest is that the equilibrium is unaffected by the change in the probability of loss. Now, we check to see if applying the original 1993 Rabin model can alone perturb the equilibrium. Below are the matrices for the $p = .25$ and $p = .75$ cases, but the utilities are calculated with the original Rabin model, using the players' expected payoffs from the normal form matrices above. In addition, we include some of the algebra used to calculate the utilities. The math is only shown for the $p = .25$ case. By convention, if the maximum possible payoff and minimum possible payoff are equal, the kindness function is just 0.

$$u_1(H, A, H, \omega, .25) = 3 + (\frac{3 - \frac{3+3}{2}}{3 - 2.5})(1 + \frac{3.875 - \frac{3.875+2.625}{2}}{3.875 - 2.625}) = 3$$
$$u_2(A, H, A, \omega, .25) = 3.875 + (\frac{3.875 - \frac{3.875+2.625}{2}}{3.875 - 2.625})(1 + \frac{3 - \frac{3+3}{2}}{3 - 2.5}) = 4.375$$

$$u_1(H, R, H, \omega, .25) = 2.5 + (\frac{2.5 - \frac{3+2.5}{2}}{3 - 2.5})(1 + \frac{1.875 - \frac{1.875+1.875}{2}}{1.875 - 1.875}) = 2$$
$$u_2(R, H, R, \omega, .25) = 1.875$$

Where u_2 here follows from the fact that Player 1 can be neither kind nor unkind in the case that Player 2 plays R. Note that due to the structure of the game, the utilities for the outcome (L,R) are the same.

$$u_1(L, A, L, \omega, .25) = 4.5 + (\frac{4.5 - \frac{4.5+4.5}{2}}{4.5 - 2.5})(1 + \frac{2.625 - \frac{3.875+2.625}{2}}{3.875 - 2.625}) = 4.5$$
$$u_2(A, L, A, \omega, .25) = 2.625 + (\frac{2.625 - \frac{3.875+2.625}{2}}{3.875 - 2.625})(1 + \frac{4.5 - \frac{4.5+4.5}{2}}{4.5 - 2.5}) = 2.125$$

		Player 2	
		A	R
Player 1	H	$(3, \boxed{4.375})$	$(\boxed{2}, 1.875)$
	L	$(\boxed{4.5}, \boxed{2.125})$	$(\boxed{2}, 1.875)$

		Player 2	
		A	R
Player 1	H	$(3, \boxed{3.125})$	$(\boxed{2}, .625)$
	L	$(\boxed{4.5}, \boxed{.875})$	$(\boxed{2}, .625)$

Again, we see a unique pure-strategy Nash Equilibrium (L,A) in both scenarios. In this case, the original Rabin model is not sufficient to explain a change in the equilibrium as the probability of loss changes. Next, let's apply our 1PR model, supposing that $\alpha = .6$ and $\beta = .8$, values that are motivated by nothing more than the satisfactory shapes of the resulting probability weighting functions. With these parameters,

$$C_\alpha(.25, v) = 1 + (2H(-v) - 1)e^{-(-ln(p))^\alpha} = \begin{cases} 1 + .296 & v < 0 \\ 1 - .296 & v > 0 \\ 1 & v = 0 \end{cases}$$

$$C_\alpha(.75, v) = 1 + (2H(-v) - 1)e^{-(-ln(p))^\alpha} = \begin{cases} 1 + .623 & v < 0 \\ 1 - .623 & v > 0 \\ 1 & v = 0 \end{cases}$$

Using these values of the corrective function, we can rewrite the matrices with the utilities resulting from the 1PR model.

$$u_1(H, A, H, \omega, .25) = 3 + (\frac{3 - \frac{3+3}{2}}{3 - 2.5})(1 + \frac{3.875 - \frac{3.875+2.625}{2}}{3.875 - 2.625})(1 - .273) = 3$$
$$u_2(A, H, A, \omega, .25) = 3.875 + (\frac{3.875 - \frac{3.875+2.625}{2}}{3.875 - 2.625})(1 - .296)(1 + \frac{3 - \frac{3+3}{2}}{3 - 2.5}) = 4.227$$

$$u_1(H, R, H, \omega, .25) = 2.5 + (\frac{2.5 - \frac{3+2.5}{2}}{3 - 2.5})(1 + \frac{1.875 - \frac{1.875+1.875}{2}}{1.875 - 1.875})(1 - 0) = 2$$
$$u_2(R, H, R, \omega, .25) = 1.875$$

Again, u_2 follows from the fact that Player 1 cannot be kind nor unkind. Thus, the value of the kindness function is 0, which means the transformed Heaviside function is 0 as well. There is no change for (H,R) between the original

Rabin model and our model.

$$u_1(L, A, L, \omega, .25) = 4.5 + (\frac{4.5 - \frac{4.5+4.5}{2}}{4.5 - 2.5})(1 + \frac{2.625 - \frac{3.875+2.625}{2}}{3.875 - 2.625})(1 + .273) = 4.5$$
$$u_2(A, L, A, \omega, .25) = 2.625 + (\frac{2.625 - \frac{3.875+2.625}{2}}{3.875 - 2.625})(1 + .296)(1 + \frac{4.5 - \frac{4.5+4.5}{2}}{4.5 - 2.5}) = 1.977$$

		Player 2	
		A	R
Player 1	H	$(3, \boxed{4.227})$	$(\boxed{2}, 1.875)$
	L	$(\boxed{4.5}, \boxed{1.977})$	$(\boxed{2}, 1.875)$

		Player 2	
		A	R
Player 1	H	$(3, \boxed{2.814})$	$(\boxed{2}, .625)$
	L	$(\boxed{4.5}, .564)$	$(\boxed{2}, \boxed{.625})$

The results demonstrate exactly what we were seeking to model: a change in the probability of loss can have major consequences on what is perceived to be kind. The changes are drastic enough to perturb the incentives and change the equilibrium. Whereas the $p = .25$ case yields the same (L,A) unique Nash Equilibrium, when p increases to $p = .75$, the Nash Equilibrium of the game changes. Player 2, given his high probability of loss, now sees a low offer by Player 1 as so unkind that he is willing to reject the offer outright before even hearing it.

5 Experimental Design

Recall the three notions that must eventually be validated through our experimentation:

a) The perceived kindness of a given outcome produced for one player by the actions of another player decrease as the probability of loss increases.

b) In circumstances in which both players are subject to a probability of loss, what is considered kind is highly reference dependent.

c) There is likely asymmetry in how each player distorts the probability space depending on whether it is the other player or himself that is experiencing the loss.

Our experimentation will initially investigate the one player case in order to determine when (a). This is in an effort to determine whether the substantive changes in equilibria demonstrated in the section four example are demonstrated in empirical behavior. Utilizing the one player equilibrium as motivation for the experiment, we will investigate whether deviations from this equilibrium occur across a range of probabilities of loss. If deviations from

this equilibrium exist, we will complete similar experimentation for the case where both players have a probability of loss; however, if we do not observe significant deviations from the above patterns of behavior, it will be unnecessary to complete experimentation for the two player case.

5.1 1PR Experimental Methods

Subjects will be told they are entering a game with one other player. The subjects will be assigned the role of Player 2, while they will be playing against Player 1. The subjects will informed that both players will be operating under the following, identical rules of engagement:

1. There exists a cumulative pool of 7.5 USD.
2. Player 1 and Player 2 will each be given an initial capital endowment of 2.5 USD from the cumulative pool.
3. Player 2 will be subject to losing the 2.5 USD with probability p, s.t. p is an element of P, where $P = [0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]$. The case where $p = 0$ represents the control, in the form of the original Rabin Fairness Model. Player 1 is not subject to losing his endowment.
4. Player 1 and Player 2 know the value of p.
5. Player 1 must decide how much of the remaining money in the cumulative pool (2.5 of 7.5 USD) to distribute between himself and Player 2 from the following options: high ($2) and low ($.75). Player 2 is aware of the restriction on Player 1's decision.
6. Player 1 will propose a value to Player 2 from the choice of high/low, which Player 2 can accept or reject.
7. If Player 2 accepts, the money is distributed according to Player 1's proposition and the probability of loss, p, is applied to Player 2 for his final payout.
8. If Player 2 rejects, Player 1 keeps the initial 2.5 USD, Player 2 keeps the 2.5 USD with probability p, and the remaining 2.5 USD is not distributed.

The following dependent variables will be recorded about their interaction:

1. Player 2's rejection rate across the defined probability space given both high and low offers.
2. The modal amount offered by Player 1 (high or low).

We suspect that significant learning will take place, and that a repeated game structure might muddy the relevance of probability as a reference point. However, given the possibility of serious heterogeniety, each subject will play the game multiple times for different probabilities of loss, as it may not be reliable to compare one subject's rejection rate at one probability with another subject's at a different probability.

While explicit parameter estimation presents a unique challenge for our model, a more reasonable analysis of the experimental results will provide insights for comparative statics. Specifically, we will determine how our endogenous variable, namely Player 2's rejection rate, responds to changes in Player 2's probability of loss.

The mathematical example of 1PR in section four identifies a unique Nash equilibrium point such that Player 2 accepts a low offer; that equilibrium is subject to change under an increase in the probability of loss. Our experimentation will seek to uncover whether this equilibrium holds for all values of p, which we do not expect to be the case based on our proposed model. Thus, upon experimentation of the 1PR case, we expect differential behavior across the probability space. We would seek to identify at what probabilities of loss Player 2's conception of kindness changes such that the equilibrium changes.

By observing Player 2's rejection rate, we will observe convergence towards equilibria for various probability and offerings combinations. For example, if we observed after n trials that out of all subject pairings at p = .1, 80% of subjects rejected low offers, we would suspect that for this probability, the kindness perceived by Player 2 was more extreme than could be explained by the original Rabin model with expected utility. After performing this analysis across the probability space, we will identify the implied equilibrium at each probability and compare this to the changes in equilibrium generated in the mathematical example from section four.

We might then apply a basic hypothesis test for the difference of proportions when comparing different probabilities of loss. Should we find that the rejection rate between two probabilities of loss is statistically significant, we would conclude that the 1993 Rabin Model is insufficient to model social interaction under risk, and that our proposed model presents a viable framework for understanding the dynamic behavior across the probability space.

6 Conclusions and Concerns

What we propose is an update to the 1993 Rabin Fairness model that incorporates behavior under risk. We have presented three intuitive notions that guide the model we outline in this proposal. Our model takes into account these three notions, and accounts for distortions in the probability space using Prelec's 1998 weighting function. Then using a mathematical example, we demonstrate that our model has the power to explain changes in equilibrium outcomes that are not explicable under the original Rabin Fairness model with expected utility. Finally, we suggest an experiment that will determine whether or not the non-standard behavior outlined in our 1PR model formulation exists in real behavior. Depending on the results of the 1PR experimentation, we would extend our experimental design to include another setup in which Rabin fails to explain a shift in equilibrium behavior when two players are subject to some risk of loss.

Despite the demonstrated plausibility of the model, there are some glaring concerns regarding its predictive power. The primary limitation of the model follows from our use of Rabin 1993 as a foundation. The model is highly sensitive to the magnitude of utilities. Since the kindness function is normalized to be in $[-1, 1]$, the product of the kindness terms in Rabin are restricted to [-2,2]. For this reason, it is difficult to perturb utilities which are of too high a magnitude, and similarly, small utilities produce highly volatile equilibria, which is not a result we deem realistic. This limitation extends to our model by virtue of our use of Rabin's implementation of the kindness functions. Ideally, an improved model would find a way to either normalize the payoffs or avoid normalizing kindness in a way that preserves the attractive features of the model. Another challenge to the model will likely be the overall mathematical complexity of how we formulate the corrective term. Upon introspection, the parameter values for the current functional form would be incredibly difficult to determine, which might suggest that there exist more elegant ways to model the same notions.

7 Worked Cited

1. Rabin, Matthew. "Incorporating Fairness into Game Theory and Economics." American Economic Review (1993): 1281-302. Web.

2. Koszegi, Botond, and Matthew Rabin. "Reference-Dependent Risk Attitudes." American Economic Review 97.4 (2007): 1047-073. JSTOR. Web.

3. Viscusi, W. Kip, and Ted Gayer. "Behavioral Public Choice: The Behavioral Paradox of Government Policy." SSRN Electronic Journal (n.d.): n. pag. Mercatus Center, Mar. 2015.

4. Schmidt, U. (2003). Reference dependence in cumulative prospect theory. Journal of Mathematical Psychology, 47(2), 122-131. doi:10.1016/s0022-2496(02)00015-9.

5. Kahneman, D., and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Emmitsburg, MD: National Emergency Training Center.

6. Tversky, A., and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty J Risk Uncertainty, 5(4), 297-323. doi:10.1007/bf00122574.

7. Koszegi, B., and Rabin, M. (2006). A Model of Reference-Dependent Preferences. The Quarterly Journal of Economics, 121(4), 1133-1165. doi:10.1093/qje/121.4.1133.

8. McNeill, John. "Other Regarding Preferences 2." Brown University, Providence. 7 Apr. 2016. Lecture.

9. Prelec, Drazen. "The Probability Weighting Function." Econometrica 66.3 (1998): 497. JSTOR. Web.