Quantitative Tests of the Perceived Relative Argument Model:

Reply to Guo and Regenwetter

Graham Loomes

Behavioural Science Group

Warwick Business School

University of Warwick

CV4 7AL UK.

g.loomes@warwick.ac.uk

Abstract

Guo and Regenwetter take PRAM 2010, add various auxiliary assumptions of their own about the utility of money, make assumptions about possible stochastic specifications, and test the various combined models against data from an experiment they conducted. However, their modelling assumptions are questionable and their experiment is unsatisfactory: the stimuli omit crucial information; the incentives are weak; and the task load is excessive. These shortcomings undermine the quality of the data and the study provides no new information about the scope and limitations of PRAM or its performance relative to other models of risky choice.

Acknowledgements: I thank the editor and two referees for many helpful comments. My thanks also to the UK Economic and Social Research Council for support under the Network for Integrated Behavioural Science programme, Grant Number ES/K002201/1, and the Leverhulme Trust for its support under the 'Value' Programme, RP2012-V-022.

1. Preliminary Remarks

In Loomes (2010) I tried to avoid claiming too much. PRAM is explicitly a model about binary choices involving lotteries which have, between them, no more than three payoffs[1]. Such tasks have generated much of the data from experimental studies of risky decision making, but they constitute a narrow class of decision problems and PRAM focuses exclusively upon them. So PRAM is not a general theory and no such claim was made.

What the 2010 Abstract *did* claim was that, with respect to this narrow class of decision tasks, the model "organizes more of the data than any other extant model". But it was stated in several places in the paper that PRAM did not claim to organize *all* of these data. In various sections, attention was drawn to observed patterns of response that *cannot* be accommodated by PRAM, together with a discussion of possible reasons, in terms of factors deliberately omitted from the model. So running an experiment to test PRAM as if it were a complete model of binary choice, even within this domain, is missing the point. Since all other models also only provide partial explanations, the relevant questions relate to how PRAM compares with the best alternatives available and which of PRAM's omissions are most important. GR do not engage with these questions at all.

2. Stochastic Specifications

PRAM 2010 was presented as a deterministic model. It was made clear that this was purely for expositional simplicity and that there was a need to extend the model in due course to allow for the stochastic element in experimental data (see Loomes, 2010, p.902 and footnote 2). However, the question of how to achieve such an extension was left open.

GR consider two types of stochastic specification. One supposes there is a fixed preference to which 'error' is added; call this an E model. The other – the random preference (RP) approach – supposes it is as if the parameters of the relevant functions are drawn at random from some underlying distribution(s). Reading GR's Comment, one might think that these are mutually exclusive accounts. But they are not.

Different variants of E model have been widely used, often with little discussion of what the error term represents. An exception is Birnbaum (2011, p.681), which lists a number of possible sources of error: some forms of presentation may make it more difficult to absorb the necessary information; there may also be mistakes in reading and/or in calculation; and fatigue may set in and

[1] To the extent that Certainty Equivalents and Probability Equivalents for two-payoff lotteries can be modelled in binary choice terms, it has implications for these types of decisions too.

concentration may lapse. Such factors are quite independent of the individual's underlying preferences, so we might think of them as constituting *extraneous* variability in people's choices.

The RP approach attributes variability to fluctuations in the individual's feelings about the relative values of different payoffs or probabilities, perhaps due to some randomness in neuronal activity or changeability in mood or attitude. This kind of variability is part of her personal decision-making apparatus and is *intrinsic*. Instead of having some fixed preference, it is more appropriate to think of her as having some probability of judging one or other option to be better at any particular moment, with that probability derived from some distribution of underlying preference functions.

Since intrinsic and extraneous variability are different concepts, they can easily co-exist. However, if the stochastic character of individuals' responses to experimental tasks is due to a combination of both intrinsic and extraneous variability, then GR's tests are compromised. To see why, consider an individual whose observed variability is due to a mix of RP and E. Suppose his RP distribution of underlying preference functions is such that, in the absence of any extraneous error, he would consider option A better than option B 80% of the time. However, if extraneous noise causes him to 'tremble' with probability 0.1, we will observe him choosing A on 74% of occasions and B on the other 26%[2]. Those of GR's tests which are based on the incorrect assumption of fixed-preference-plus-extraneous-noise will wrongly interpret this as if he has a fixed preference for A and a tremble probability of 0.26: that is, such a test would judge his tremble to be more than twice its true size. On the other hand, those tests which are based on the incorrect assumption of random-preference-plus-NO-extraneous-noise will infer wrongly that his underlying preference functions entail a core probability of choosing A of 0.74 rather than 0.80.

If GR's tests are mis-specified in these respects, then when a model is deemed to have failed, it is hard to tell whether it is some key assumption of the model that has failed or whether it is due to an incorrect assumption built into the test.

It is straightforward to show that if RP and E co-exist, the expected frequency with which an individual will choose his more-likely-preferred option will *always* be less than the RP-only true probability of preferring it; and the 'tremble' coefficient inferred from observation will *always* tend to be greater than its true value[3]. This may help to explain why GR's rejection rates leap when the error parameter τ is limited to 0.25: as the example shows, the true tremble rate may be 0.1, but

[2] The arithmetic for a representative set of 100 choices is as follows. On the 80 occasions where a randomly drawn function implies that A is better, the decision maker trembles with probability 0.1 and actually picks B 8 times and A 72 times. On the 20 occasions where RP favors B, his 0.1 tremble leads to A being chosen twice and B 18 times. Total observed A choices: 72 + 2 = 74. Total B choices: 8 + 18 = 26.

[3] Proof available from the author on request.

when combined with RP variability, it is wrongly estimated by the E-only model to be greater than 0.25. So those rejection rates may in part be a reflection of the mis-specified nature of GR's tests.

That problem is compounded when GR start to make assumptions about the function c(.) which maps money payoffs x_i into subjective values $c_i = c(x_i)$. Loomes (2010) gave little explicit guidance about how to specify c(.). Although many people may behave as if c(.) is concave, this is not crucial: all that really matters is that c(.) is increasing in x, so that if $x_2 > x_1$, then $c(x_2) > c(x_1)$. Because the main results in Loomes (2010) do not depend on the particular shape of c(.), it was stated in footnote 12 that it was simpler to derive some of those results on the basis of a linear c(.). But that was explicitly just a matter of expositional convenience.

It would appear that GR took those analytically convenient simplifying assumptions too literally. They focus initial attention on two specifications: one that they label **Id**, where $c_i = x_i$; and another labelled **PwrA** where $c_i = x_i^{\rho}$ with ρ constrained to lie between 0.01 and 1. Under these particularly restrictive assumptions, they can generate some specific predictions, as listed in their Table 1. However, there is no rationale for those restrictions: it is entirely legitimate to allow c(.) to be convex – or indeed, to have some concave and some convex ranges; and it is perfectly permissible to allow the distribution of c(.) functions that underlie an RP model to consist of some mix of concave, linear and convex functions[4]. However, once we move to the less restrictive **Pwr** or **Quad** models, the bulk of GR's predictions evaporate: only four qualitative predictions survive in GR's Table 1 – and these are the ones adapted from the experiment in Loomes (2010).

The truth is that *none* of the seven particular specifications of c(.) listed in GR's Table 1 is intrinsic to PRAM: they are all auxiliary assumptions and could just as well be applied to EU, RD or a number of other models[5]. All of the models considered by GR are actually combinations of the distinctive features of PRAM (epitomised by the parameters α and δ) together with different specifications of c(.). Since the assumptions relating to α and δ are held constant across the seven variants, the considerable differences in the performance of the various models reflect little more

[4] Famously, Friedman and Savage (1948) proposed a function that was concave and then became convex; and although Markowitz (1952) was sceptical about their particular proposal, his alternative also had concave and convex ranges. Hey and Orme (1994) allowed the data great freedom and their estimates – see their Table 4 – suggested that although strictly concave functions were predominant, there were a small minority exhibiting strict convexity and a much larger minority consistent with functions that were concave for lower payoffs and convex for higher payoffs. In the Loomes et al. (2002) study cited earlier, both the EU and the RD counterparts of c(.) were allowed to span the range of convex, linear and concave functions. Restricting c(.) to be concave or linear has no basis in history, theory or evidence.

[5] The experimental design in Loomes (2010) required *no assumptions whatsoever* about c(.) except that it was increasing in x.

than the relative performance of the auxiliary assumptions about c(.)[6]. They have nothing useful to tell us about PRAM *per se*. Of course, GR are free to explore as many functional forms of c(.) as they like; but they are quite wrong to present rejections of any particular form of c(.) as if it is somehow a rejection of what is distinctive and central to PRAM.

3. Issues Concerning Stimuli, Incentives, Task Load and the Quality of GR's Data

A major advantage, in principle, of using experiments to examine decision theories is that they can instantiate key features of any theory under scrutiny and can try to keep extraneous confounds to a minimum. In practice, much depends on the quality of the experiment. How well – or poorly – does GR's experimental environment implement PRAM?

As Figure 1 in Loomes (2010) shows, PRAM assumes that decision makers compare the higher payoffs from both options with each other, and likewise for the lower payoffs. Knowing the probabilities involved, the individual is then assumed to act as if she computes the relative chances of each option doing better than the other, compared with the relative differences in the subjective values of the payoffs.

To stay close to the theory, the experiment reported in Loomes (2010) mimicked the format in Figure 2 of that paper. The payoffs were lined up appropriately, together with their respective *and explicitly stated* probabilities, making it easy to see the relevant differences. If participants were inclined to make the judgments proposed by PRAM, they could do so. If they were inclined to act according to EU or RD or any other model, it was just as easy to do that.

However, GR made it difficult to undertake those judgments. Contrary to the great majority of risky choice experiments, they omitted all explicit information about the probabilities: instead, respondents had to try to figure them out by visual inspection of the segments of 'wheels of chance'.

Before going further, readers might like to look at Figure 2 in GR's Comment and try to judge the probabilities of the different payoffs in each of the two wheels there; or alternatively (as far as PRAM is concerned), they might try to judge how much higher the probability of $15 is in the right-

[6] It is potentially misleading to claim in their Abstract that their conclusion about model performance "is robust across 7 different utility functions for money" and in in their Conclusion that "the error model analyses of PRAM consistently show poor model performance, regardless of . . . which of seven utility functions of money we use . . .". The rejection rate of the frequentist modal choice **Id** model is 62.8%, while the comparable rejection rate for the **Pwr** model is just one quarter of that, namely 15.7%; and **Quad** 'fits' more than 88% of cases (see the top panel of GR's Table 4). Since there is no comparison of this rate with any other model (e.g. EU or RD using the same set of c(.) functions), it is hard to draw any conclusion about how 'poor' an 88% fit rate is. However, suggesting that rejection rates that vary between 62.8% and 11.6% are 'consistent' certainly seems to be a questionable use of the word.

hand wheel than in the left-hand wheel (this is b_R in PRAM – see Expression (2) of the Comment), and also how much higher the probability of $5 is in the right-hand wheel than in the left-hand wheel (this is b_S in that Expression).

GR's tests are conducted on the assumption that respondents act as if they perceive all probabilities precisely and accurately and choose on the basis of the true b_S and b_R. Readers who made their own estimates for GR's Figure 2 may now like to compare them with the probabilities for Pair 14 in GR's Table 2, from which we derive $b_S = 0.07$ (that is, the difference between 0.44 and 0.37) and $b_R = 0.40$ (the difference between 0.56 and 0.16). This gives a b_S:b_R ratio of 1:5.71. Quite small misjudgements of the probabilities and the differences between them can significantly affect the perceived ratio (and hence the implications for PRAM-based behavior): someone who judges b_R correctly but judges b_S to be 0.05 will thereby produce a ratio of 1:8, while judging b_S to be 0.10 would give a ratio of 1:4. Variations of this kind can have substantial effects upon the probabilities of choosing S or R as compared with the true ratio of 1:5.71 and can thereby introduce a considerable amount of additional error into patterns of response.

This threatens to undermine the quality of GR's data and their subsequent analysis. They could simply have put the probabilities as well as the payoffs on their displays, but they chose not to do so. Instead, they adapted a format used by Tversky (1969, his Figure 1) which Tversky chose, *not* in order to make the probabilities transparent, but in order to deliberately obscure them, as he explained on his p.33. Obscuring the probabilities may have served Tversky's purposes in that 1969 study, but it is a poor design for providing a fair test of any theory constructed on the basis that the decision maker knows the probabilities.

There are further factors that may have aggravated this problem. GR's incentive mechanism involved paying participants on the basis of one decision chosen at random from among their 1,600 responses. On this basis, the chance that any one decision will really matter is miniscule. This provides a very low incentive to invest much care and effort in any choice. Moreover, participants had very little time to do so. GR gave a guideline of 1.5 hours to complete each session. Working non-stop through 1,600 choices in 90 minutes involves spending just over 3 seconds on each choice. There must be doubts about the accuracy with which anyone can estimate the probabilities *and* integrate them with payoffs *and* produce a reliable statement of preference under these conditions. As participants worked through this demanding repetitive task, it would not be surprising if fatigue and lapses of concentration produced yet more extraneous noise.

How might we explore the extent of such noise in this dataset? Ideally, we should identify some choice where the 'true' preference is known and where, in the absence of extraneous error, option A would be chosen by everyone with probability 1, even in an RP model, while option B would never be chosen. If we then observe B being chosen with frequency ε, this ε may provide some broad indication of the amount of extraneous error (and also the extent to which it varies between individuals).

Fortunately, one of GR's Pairs effectively fits this bill. Figure 1 reproduces GR's Pair 20. Visual inspection suggests that both options offer the same chance of \$5. The rest of the right-hand (RH) wheel gives \$15 while the left-hand (LH) wheel gives a rather smaller chance of \$15 and a substantial chance of getting \$10 instead of \$15. In effect, RH stochastically dominates LH and there is no good reason to do anything other than choose RH[7].

FIGURE 1 ABOUT HERE

When experiments present respondents with such choices in a form that makes the dominance relationship easy to see, and give them time to deliberate, the rate of violation is often very low. For example, in Loomes and Sugden (1998) five 'dominance' questions were included in the experiment, scattered among 40 other questions, and the rate of violation was less than 1.5% – a rate which is fairly typical of such studies.

There is one important exception. Birnbaum's TAX model (see Birnbaum, 2008, especially pages 473-4) *does* predict substantial violations of stochastic dominance if a pair of lotteries conforms to his "special recipe". But GR's Pair 20 does *not* conform to that special recipe; so TAX, like every other mainstream model, predicts that the RH option will be preferred with probability 1 under regular assumptions[8]. Thus the frequency of choosing the inferior LH option provides a measure of the propensity to tremble, ε, that reflects extraneous error. As we shall see, that single simple measure turns out to be highly diagnostic.

[7] The 0.003 difference in the probabilities of \$5 means that, strictly speaking, RH does not dominate LH; but this difference is arguably too tiny to be detected by anyone with normal eyesight and powers of discrimination.

[8] One can enter the probabilities and payoffs into an online TAX calculator and confirm this result under what Birnbaum regards as the default parameters: go to http://psych.fullerton.edu/mbirnbaum/taxcalculator.htm To be sure, I asked Dr Birnbaum (email correspondence) whether there were circumstances under TAX such that the LH option might be chosen. He thought this result could be produced if the \$10 payoff was treated by respondents as being worth \$12.50 or more while the \$5 and \$15 payoffs were taken as they are – that is, if there was a very high degree of concavity of the c(.) function over that range. If this were the case, we should expect to see individuals who pick LH most often also displaying consistently greater risk aversion over the other Pairs; but, as we shall see shortly, they do not.

For the 54 individuals who participated in both sessions, the mean rate of choosing the inferior LH option in the first session was 27.6%, rising to 32.3% in the second session – that is, a rate *ten or fifteen times higher than usually found*. This high average was not driven by a minority of extreme responses: the *median* rate was 30%; and 77.8% of participants chose LH on at least 10% of occasions. If this level of extraneous error was occurring in this 'easy' choice, we might wonder whether it affected all of the other responses, rendering the whole dataset highly unreliable.

To investigate this, let us divide the 54 individuals who participated in both sessions into three subsamples of approximately equal size, differentiated by the frequencies with which they chose LH. The subsamples are as follows:

LowE consists of 19 individuals who chose LH 7 times or fewer out of the total of 40 times they were presented with Pair 20 in the two sessions – that is, an ε of less than 20%. The mean ε rate for this subsample was 7% and the median rate was 5%.

MidE consists of 17 individuals who chose LH between 8 and 15 times – an ε rate between 20% and 40%. For them, the mean rate was 29% and the median was 30%.

HighE comprises 18 individuals who chose LH 16 times or more – an ε rate of at least 40%, with a mean of 55.15% and a median of 51.25%.

First, consider Pairs 1–6. These questions had simpler displays (with certainties in three cases and probabilities that are 'round' decimals in other cases), with one payoff being zero (making calculations somewhat easier). In all six pairs, the riskier (R) options offer expected values between 10% and 60% higher than their safer (S) counterparts. Someone who is only moderately risk averse and who is paying attention to the stimuli could be expected to choose the R option quite often.

The top panel of Table 1 shows the mean and median numbers of S choices per person, with percentages in brackets. For the LowE group, whose data are shown in the second and third columns, there is clear evidence that the R options were chosen much more often: the majority of members of LowE chose the S option on fewer than 15% of occasions.

If members of MidE were similar to LowE in terms of underlying preferences but were more susceptible to extraneous noise, we should expect them to be much less discriminating. And so it appears: MidE's rates for choosing S are shown in the fourth and fifth columns, and both are in the vicinity of 50%. Likewise for the HighE group – see the sixth and seventh columns.

TABLE 1 ABOUT HERE

Of course, it might be argued that for some reason the members of those last two groups happened to be more risk averse than the members of LowE and were therefore more inclined to choose S. The data from Pairs 7–12 show that this is not the case.

Pairs 7–12 are the 'mirror-images' of Pairs 1–6 in the sense that the probabilities of 0 and $20 are interchanged. So all six R options are unambiguously less attractive relative to S than in Pairs 1–6; and in all six of Pairs 7–12, S is not only the safer option but also offers a higher expected value[9]. Everyone who is paying attention and who is reasonably responsive to these differences should choose S (much) more often in Pairs 7–12 than in Pairs 1–6. So if the main difference between the three groups relates to the degree of extraneous noise, we should expect to observe the least responsiveness from HighE and the greatest responsiveness from LowE.

This is what we find in the second panel of Table 1. There is *some* responsiveness in the HighE group, with the median rate of choosing S increasing 7% and the mean rate going up less than 10%. There is greater responsiveness among members of MidE, with mean and median rates about 28% and 38% higher than for Pairs 1–6. For LowE the change is much more striking: the mean rate is 52.5% higher and the median goes up 85%. Hence the ordering between groups is completely reversed: now LowE has the highest proportion of S choices while HighE exhibits the lowest proportion, showing how substantial variability in the amounts of extraneous noise can interact with different sets of choices to produce very different conclusions about relative risk aversion.

The third panel of Table 1 considers another six Pairs. Just as in Pairs 1–6, the R options offered higher expected values (on average, 12.4% higher) than their S alternatives[10]. Again, the contrast with Pairs 7–12 is sharpest for LowE, with mean and median differences between 60% and 80%; for MidE those differences are between 38% and 49%; while for HighE they are just 10% and 8%. So here too, sensitivity to lottery parameters was much poorer at intermediate levels of noise and was largely swamped when noise levels were high.

So much noise among so many participants could explain why GR's tables show huge rejection rates when τ is set at 0.25: even in the two least restrictive cases of **Pwr** and **Quad**, the frequentist rejection rates leap from 15.7% and 11.6% when $\tau = 0.5$ to 79.3% and 77.7% respectively when $\tau = 0.25$. One interpretation is that this leap testifies to the widespread low quality of the data.

[9] Pairs 3 and 9 illustrate the contrast. In Pair 3, the R option offers a 0.8 chance of $20 and a 0.2 chance of 0, while in Pair 9 the probabilities of $20 and 0 are reversed, with R offering a 0.2 chance of $20 and a 0.8 chance of 0. For Pair 3, R's expected value was 60% higher than S's; for Pair 9 it was 60% lower.

[10] Pair 18 is different from the others in this respect since its S option has a higher expected value. It is therefore more like Pairs 7–12.

The prime suspects are the ‘wheel of chance’ formats that obscure crucial probability information, with this source of error amplified by heavy task load and miniscule incentives.

References

Birnbaum, M. (2008). New paradoxes of risky decision making. *Psychological Review*, **115**, 463–501.

Birnbaum, M. (2011). Testing mixture models of transitive preference: comment on Regenwetter, Dana, and Davis-Stober. *Psychological Review*, **118**, 675–83.

Friedman, M. and Savage, L. (1948). The utility analysis of choices involving risk. *Journal of Political Economy*, **56**, 279-304.

Guo, Y. and Regenwetter, M. (2014). Quantitative tests of the perceived relative argument model: comment on Loomes (2010). *Psychological Review*, **121**, xxx-xxx.

Hey, J. and Orme, C. (1994). Investigating generalizations of expected utility theory using experimental data. *Econometrica*, **62**, 1291-1326.

Loomes, G. (2010). Modeling choice and valuation in decision experiments. *Psychological Review*, **117**, 902-24.

Loomes, G. and Sugden, R. (1998). Testing different stochastic specifications of risky choice. *Economica*, **65**, 581-98.

Markowitz, H. (1952). The utility of wealth. *Journal of Political Economy*, **60**, 151-8.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, **76**, 31-48.

FIGURE 1: GR's Pair 20

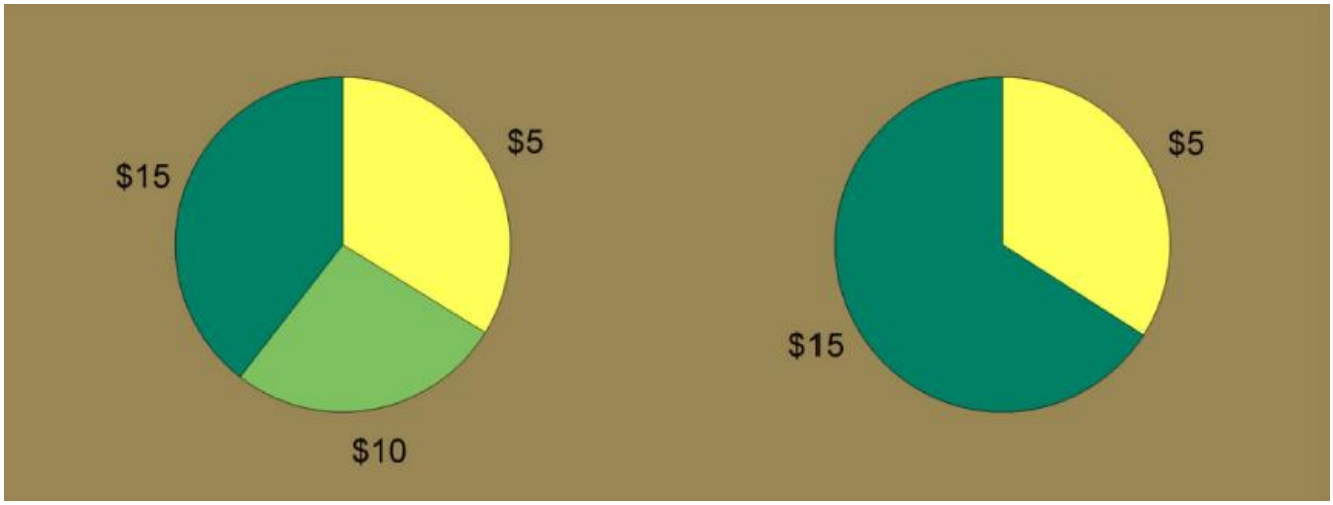

TABLE 1:

	LowE (n=19)		**MidE** (n=17)		**HighE** (n=18)	
Pairs	Mean	Median	Mean	Median	Mean	Median
1–6	73.05 (30.4%)	32 (13.3%)	122.41 (51%)	111 (46.3%)	139.11 (57.9%)	137 (57.1%)
7–12	199.05 (82.9%)	236 (98.3%)	190.18 (79.2%)	202 (84.2%)	161.17 (67.2%)	154 (64.2%)
13–17, 19	55.37 (23.1%)	50 (20.8%)	99.65 (41.5%)	86 (35.8%)	138.22 (57.6%)	135.5 (56.5%)