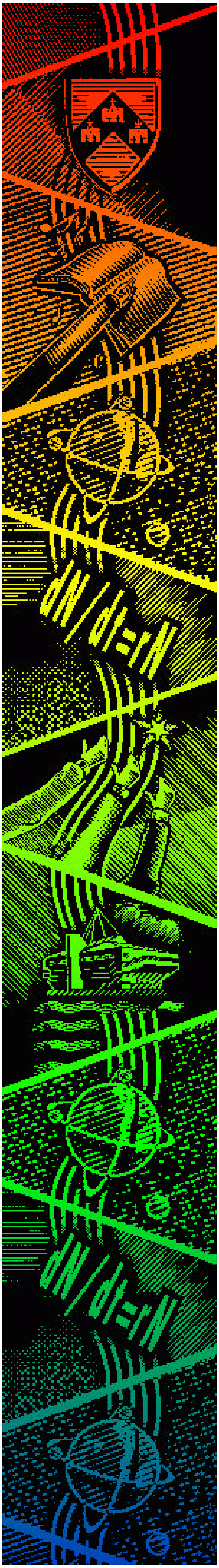

THE UNIVERSITY of York

Discussion Papers in Economics

No. 1999/28

Does Repetition Improve Consistency?

by

John Hey

Department of Economics and Related Studies
University of York
Heslington
York, YO10 5DD

Does Repetition Improve Consistency?

John D. Hey[*]
Universities of York and Bari

September 30, 1999

Abstract

Much experimental effort has been expended in attempts to establish the relative superiority of Expected Utility theory and the many recently-developed alternatives as descriptions of the behaviour of subjects in risky choice decision problems. The cumulative evidence shows clearly that there is a great deal of noise in the experimental data, which makes it difficult to identify the 'best' description of such behaviour. This paper reports on an experiment which seeks to determine whether such noise is relatively transitory and decays with experience and repetition, and thus whether a clearly 'best' theory emerges as a result of such repetition. We find that for some subjects this does indeed appear to be the case, while for other subjects the noise remains high and the identification of the underlying preference function remains difficult.

1 Introduction

Early experimental work suggesting the inadequacy of Expected Utility theory as a description of behaviour under risk has stimulated, over the past decade, a large number of alternative theories of such behaviour. This, in turn, has stimulated new experimental activity investigating the relative superiority of these new theories. Interestingly, no clear concensus has emerged. One main reason for this is that the experimental evidence clearly indicates that there is a large amount of noise in subjects' behaviour in such experiments - which makes accurate testing and estimation difficult. The existence of the noise is itself interesting as most of the contending theories have deterministic predictions. This necessarily suggests that none of the theories is empirically valid, at least insofar as the existing empirical evidence is concered. A crucial point, therefore, is whether the empirically-observed

[*] The research reported here was funded under a grant (R000 23 6636) from the Economic and Social Research Council of the UK "Experimental Investigations of Errors in Decision Making". I am grateful to Marie Bissey and Vittoria Levati, Reseach Fellows on this grant, for assistance in running the experiments and for discussions concerning the analysis of the data.

noise is a *permanent* feature of subjects' behaviour - and thus 'ought' to be integrated somehow into the theories - or whether the noise is simply a consequence of the experimental design. In particular, it has been argued that the noise is a consequence of the fact that most experiments involve inexperienced subjects in unfamiliar situations tackling a one-off decision problem. Consequently it may be the case that the noise is transitory and will disappear in an experiment in which subjects have the opportunity to learn and understand the nature of the problem. If so, it might be expected that the noise will tend to disappear with experience, thus facilitating the better identification of the 'best' description of the underlying true preference function. This paper reports on an experiment designed to explore this line of reasoning.

2 The Experiments

We wanted to give the subjects plenty of opportunity to experience the decision task, so we recruited a set of subjects who would repeat the experiment a total of 5 times. In the past subjects have usually been exposed to the experiment on just the one occasion, though there have been a number of experiments in which some element of repetion has been involved, either within one experimental session (the same questions repeated within the one experimental setting) or in two (but not to our knowledge more) experimental settings; indeed, we used this technique ourselves earlier (Hey and Orme, 1994) with the same questions (though in a randomised order and presentation) on two separate sessions, separated by a period of several days. In the experiments reported in this paper, we repeated the same set of questions in a total of five sessions. As before, we randomised the order in which the questions appeared (so that each subject in each session receieved the questions in a different order).

The experiment was computerised, with subjects arriving individually at pre-booked times. For all subjects, the five sessions were separated by at least two days, to give them time to reflect on the experiment and on the way that they were responding to it. The experiment was computerised and the subjects performed the experiment individually at individual computer terminals in the EXEC laboratory.

On each occasion the subjects were presented with 100 pairwise risky-choice questions, portrayed on the computer screen in the form of segmented circles. They were asked to indicated which of the pair they preferred[1]. To avoid misperception of probabilities, all the questions involved probabilities that were multiples of one-eighth, and the subjects were informed of this. This circles form of presentation is one that we have used several times before and we feel that subjects have no difficulty in understanding the probabilistic implications (though we were careful to avoid the use of the word 'probability' in the instructions[2]. The subjects were informed that, after the completion of all five experimental sessions, one of the 500 questions (100 on each of the 5 sessions) would be chosen at random and their choice would be played out (using a (circular) 'roulette wheel') and they would be paid accordingly. This 'random lottery incentive mechanism' is one that is frequently used in such experiments. If subjects examine each pairwise choice question independently of the others (which implies an assumption of separability) then this payment mechanism gives them the appropriate incentive to reveal their true preferences on each question. Of course, it could be argued that subjects consider the experiment as a single decision problem - that of choosing a complete set of 500 answers to the 500 questions - and choose that set of answers which is optimal for them given their preference functional. This is the argument of (Karni and Safra, 1987) and of (Holt, 1986). If subjects did this there is no obvious reason why they should give the same answers to the same question, but it requires a level of sophistication that seems far beyond what subjects might be capable of. Indeed it requires them knowing in advance all the 500 questions (which they could not) and then considering all 2^{500} possible sets of answers and then choosing that one that was optimal for them. This seems far beyond the computational powers of any human subject - even

[1] In these experiments, in contrast to earlier experiments we have conducted, we did not give the subjects the opportunity to indicate indifference. This does not affect the value of the experiment to the subjects, since if they are truly indifferent it does not matter how they respond, given the incentive mechanism (described below), but it simplifies our subsequent data analysis: if subjects are given the opportunity to express indifference, and take advantage of this opportunity, it is not obvious how one should treat such responses - given the nature of the stochastic structure we assume, the chance of them being exactly indifferent is zero, which implies a log-likelihood of minus infinity, which in turn makes subsequent estimation difficult.

[2] We appreciate that other experimentalists use other methods for presenting risky choices and we would not argue that ours is the best. Indeed, for certain types of experimental investigations, particularly those relating to Regret theory, our presentation might well not be appropriate. But in the context of this experimental investigation, we feel that the circles method is perhaps the simplest and most appropriate, particularly when considered in conjunction with our payment mechanism (see below).

assuming that they knew all 500 questions. Like other experimentalists, we shall ignore this possibility, citing the work of (Cubitt et al., 1998) which indicates that the Random Lottery Incentive mechanism does indeed work as we require it to work.

The 100 questions in each of the five experimental sessions were the same 100 questions, though the order was randomised, and the left/right positioning of the circles was also randomised. The 100 questions were composed of four different sets, each consisting of 25 questions. Over all 100 questions the possible outcomes in the questions were -£25, £25, £75 and £125, these amounts being chosen to make the incentive offered by the experiment appropriate given the length of time necessary to complete it. The 4 sets of questions each involved just three of these four possible outcomes: the first set restricted to £25, £75 and £125; the second set to -£25, £75 and £125; the third set to -£25, £25 and £125; and the fourth set to -£25, £25 and £75. Accordingly, each pairwise choice question involved a choice between two gambles which between them involved at most three outcomes. Note that one of these four outcomes (-£25) involved a monetary *loss*. We felt that this would increase the incentive power of the experiment. At the same time, given that we did not want any subject to experience a real monetary loss, we gave all subjects a participation fee of £25 for participating in all 5 sessions of the experiment. This meant that the real monetary earnings of each subject would be one of the four possible amounts of money: £0, £50, £100 and £150. This implied that our preference functionals were effectively fitted (in payoff space) at the four values £W-25, £W+25, £W+75 and £W+125 where £W was the wealth of the subject at the time of participating in the experiment. Clearly W could vary during the course of the experiment, but we could regard the appropriate value as the value of the subjects' wealth at the final (fifth) session of the experiment - when they received the payoff. They knew in advance when this would be. The set of questions is given in Table 1.

We should briefly comment on the choice of these questions. In doing so, it will be useful to refer to the expository device now known as the Marschak-Machina Triangle, which allows the representation of choice over risky choices involving at most three outcomes (as in our experiment). If we denote the three outcomes by x_1, x_2 and x_3 and if we label these so that

x_1 is the least preferred outcome and x_3 the most preferred, then any risky choice involving these three outcomes can be represented by two numbers: p_1 and p_3, the probabilities respectively of the worst and best outcomes. (p_2 is, of course, given by $1 - p_1 - p_3$.) We put p_1 on the horizontal axis and p_3 on the vertical. The set of all possible risky choices involving the three outcomes x_1, x_2 and x_3 is represented by the triangle with vertices at the origin, at (0,1) and at (1,0). Now envisage a rectangular grid imposed over this triangle, with a grid difference of 1/8. This defines a set of 45 points, each representing a risky choice with probabilities multiples of 1/8. Any pair of such points represents a pairwise choice problem. If the pair is such that one point is above and/or to the left of the other then the first of these *dominates* the second in the sense of first degree stochastic dominance: as such all subjects should prefer the risky choice above and/or to the left - whatever the form of their preference functional (over risky choices). It might be argued that including such pairs is not necessary in the sense that their inclusion does not allow discrimination *between* the various competing preference functionals: if a subject chooses the dominating choice then this behaviour is consistent with all preference functionals; if a subject chooses the dominated choice then this behaviour is inconsistent with all preference functionals. Indeed, many experiments have deliberately excluded such pairs. However, six (out of the 100) of the pairwise choice questions used in this experiment were of this form. We wanted to test whether behaviour violated dominance. The six questions are questions 24, 25, 49, 50, 75 and 100 in Table 1. In the past, experimentalists have usually observed that subjects' behaviour does not violate dominance when that dominance is 'obvious'. We wanted to see whether that was also true in our experiment. We shall comment on this later.

The remaining questions all were such that one of the two points lay strictly above and to the right of the other: thus whether a subject prefers one or the other depends upon his or her preference function. More particularly, a subject's preference in a particular pairwise choice question depends upon his or her local (see (Machina, 1982)) risk aversion at that point in the triangle. Given any pairwise choice, the more risk averse the subject the more likely it is that he or she prefers the (relatively) safer of the two risky choices - namely that below and to the left of the other. Now, as is well-known, in the Marschak-Machina triangle (where the outcomes, x_1, x_2 and x_3, are fixed) risk-aversion, for a subject with the Expected Utility preference functional, is constant, so preference between any two choices in a pairwise choice problem for such a subject should depend solely on that subject's level of risk-aversion and the slope of the line joining the two points in the triangle. Crucially, the position of the two points should be irrelevant: for an Expected Utility maximiser just the *slope* of the line joining the two points is relevant.

Another way of putting the same point is to note that the indifference curves of an individual in the Triangle depend upon the preference functional. For an Expected Utility maximiser we have the important result that the indifference curves are parallel straight lines - with a (constant) slope that depends upon the individual's attitude to risk. For other preference functionals, risk-aversion varies across the triangle, so the slope of the indifference curves vary across the Triangle. Different preference functionals are distinguished by the precise way that the slope varies across the Triangle - for example, Weighted Expected Utility theory implies that the slope decreases as one moves down and to the right - and this is the way that we can empirically distinguish between the different preference functionals. Clearly it is therefore important to be able to estimate the slope of an individual's indifference curves at as many points in the Triangle as possible.

Unfortunately, pairwise choice questions do not yield a direct estimate of the slope of the indifference curve - all they yield is upper or lower bounds on the slope. For example, suppose the slope of the line joining the two points in a pairwise choice question in s, and suppose that the subject says that he or she prefers the risky choice above and to the right. Then, assuming for the moment that there is no error in the subject's reported preferences, this observation enables us to conclude that the (average) slope of the individual's indifference curve in the region of the two points defining the pairwise choice is *less than* s. Contrariwise, if the subject reported a preference for the risky choice below and to the left, what we can conclude is the (average) slope is *more than* s. In other words, pairwise choice data is not particularly informative.

This might suggest that we should have looked for more informative data. For example, if we know that a subject is exactly indifferent between two choices, then (again ignoring any error) we would have a precise estimate of the slope of the individual's indifference curve in that region - namely the slope of the line joining the two points. Unfortunately, knowing "that a subject is exactly indifferent between two choices" is not as simple as it sounds. Indeed, even if we have been lucky enough to select a pairwise choice question in which the individual truly is indifferent, devising an incentive mechanism which motivates the subject to reveal that indifference is not easy: indeed our payment mechanism does not achieve this. Other experimentalists have used a sort of iterative procedure combined with a Becker-Degroot- Marschak payment procedure, but we have serious doubts about this - particularly once the subjects realise what is going on and tries to manipulate the iterative procedure to maximise their payoff (which is, of course, precisely what we want them to do). The great advantage of the simple exogenised pairwise choice procedure that we have used is its simplicity and clarity: subjects are precisely aware of what they have to do and what is in their best interests to do.

But, given that pairwise choice questions are not particularly informative, we need to ask a large number of questions. The more questions the better, as long as tiredness does not set in. We feel that 100 questions in one experimental setting achieves the desired trade-off between quantity of information and tiredness.

The precise questions used in the experiment were chosen as follows:

A total of 53 subjects participated in all five sessions of the experiment. Given that we treat all subjects as different and carry out our estimations *subject by subject*, the number of subjects is in some sense arbitrary, particularly as we can not claim that our set of subjects is in any way representative. But as the object of the experiment is to see if behaviour changes over the repetitions, and perhaps converges through repetition, the number of subjects that is appropriate is not well defined, though a number in excess of 50 seems appropriate if one is interested in some kind of average behaviour over a not-unrepresentative set of subjects.

3 A Descriptive Analysis of the Results

Before we begin to fit the preference functionals to the subjects' responses, we present some simple descriptive statistics which indicate the nature of the randomness in behaviour and show how it changes over the course of the five repetitions of the experiment. Recall that each of the five repetitions involved the same 100 questions. We can analyse whether the same responses were given by the subjects on all 5 repetitions. Our analysis is done subject by subject and is presented in Table 2. In this table, the rows represent the 53 subjects. The column headed '1 to 2' indicates, for each subject, the number of questions on which the subject's response differed between Repetition 1 and Repetition 2. Clearly the entry would be 0 for a subject who gave precisely the same responses, and would be 100 if the subject gave precisely contrary answers, that is, preference for Choice 1 on one of the two repetitions and preference for Choice 2 on the second of the two[3]. The other columns ('2 to 3', '3 to 4' and '4 to 5') are similarly defined. The column headed 'total of these' simply sums the previous 4 columns and is a crude measure of the overall randomness in subjects' responses. However, it may well be misleading in that the subject may simply switch preferences on a subset of questions. For example, suppose that a subject had fixed preferences for 90 of the 100 questions and on the remaining 10 questions oscillated regularly between the two choices between repetitions. Then for this subject there would be a '10' in the first four columns of Table 2 and a '40' in the penultimate column. Now consider a subject who also makes 10 changes in his or her stated preferences between each repetition but, in contrast, these 10 changes are on a different set of 10 questions on each occasion. This for this subject there would also be a '10' in the first four columns of Table 2 and a '40' in the penultimate column. However it could well be argued that this second subject is more inconsistent than the first. Accordingly, we present in the final column of this table a count of the number of questions on which the subject gave differing answers at some stage of the experiment. Again a '0' indicates no changes (fixed stated preferences) while a '100' indicates that the subject changed his or her stated preferences at least once in all 100 questions. For the first of the two examples presented above, the value in the final column would be 10, while the value for the second of the two examples would be 40. We note that the difference between the values in the two final columns gives an indication of the extent to which the variability is confined to a subset of the questions: a low value of this

[3] It should be noted that the subjects received the questions in a random order and with the left/right positioning randomised, but this table refers to the basic (unrandomised) questions as given in Table 1.

difference indicates the variability is spread across questions while a high value indicates that it is confined to a subset of the questions.

Let us begin with the penultimate column - which indicates the total number of consecutive changes in stated preferences. Across subjects this varies considerably, from a minimum of 4 (subject 46) to a maximum of 91 (subject 20), and there is a considerable dispersion in between. The maximum possible value of this variable is 400, so there is a reasonable degree of constancy for most of the subjects. If subjects were choosing totally at random, we would expect on average 50 changes from repetition to repetition, giving a total expected value of 200 for the figure in the penultimate column. The subject's responses indicate clearly (at very low levels of significance) that they were not anwering at random - though clearly there is a considerable degree of variability in their responses. The figures in the final column of Table 2 is generally somewhat lower that the figures in the penultimate column, and perhaps presents a more accurate picture of the variability in responses. The minimum value in this column is 3 (subject 46) and the maximum of 48 (subject 20). All the figures in this column are below 50, indicating that on at least half the questions subjects had fixed stated preferences. These figures cast considerable doubt on the Constant Probability error specification of Harless and Camerer.

A measure of whether the variability is resticted to a subset of the questions is given by the difference between the penultimate and final columns of the table. Generally this difference is of the same order of magnitude as the final column itself - varying from a low of 1 (subject 46) to a high of 43 (subject 20). These figures suggest that the probability of making a mistake (as measured by the extent to which subjects give different responses to the same question) is not constant across questions; this is consistent with the White Noise error specification. However, indirect evidence (and not conclusive evidence) against the White Noise error specification can be found in Table 3 which lists the violations of dominance observed in the 5 repetitions. Recall that 5 of the 100 questions involved a pair of risky choices, one of which dominated the other (in the sense of first-degree stochastic dominance). For all preference functionals (assuming only that the subjects prefer more money to less) we would expect that dominance would be respected - that is, subjects would choose the dominating prospect of the two. Choosing the dominated prospect would be a violation of dominance. Table 3 lists the numbers of such violations observed in our experiment. This table clearly shows that such violations are extremely rare; indeed on question 100 there were no violations at all. However, this table does suggest that on certain such questions violations are more likely to be observed than on others. It may be interesting to note that the structure of these 6 questions was as follows - in terms of the relative positioning of the two points in the Marschak-Machina triangle:

- question 24: one point to the left of the other by a distance 1/8
- question 25: one point vertically above the other by a distance 1/8
- question 49: one point to the left of the other by a distance 1/8
- question 50: one point vertically above the other by a distance 1/8
- question 75: one point to the north-west of the other by a vertical and horizontal distance 1/8
- question 100: one point to the north-west of the other by a vertical and horizontal distance 1/8

It will be noticed that in Questions 75 and 100 the dominance is 'more obvious' (and indeed stronger) than in the other questions. Perhaps it is not suprising that there are less violations with these two questions. However there is not obvious reason why violations of dominance are greatest with Question 25 - particularly as its structure is almost identical to that of Question 50. One possibility is that Question 25 involves the three positive outcomes (£25, £75 and £125) while Question 50 involves a negative outcome (-£25, £75 and £125). It could be argued that subjects tend to be more careful in answering questions which involve losses.

Returning to Table 2 we can now examine how the inconsistency rate varies across the 5 repetitions. There are clearly (at least) two different kinds of subjects: those whose inconsistency rates decline markedly over the five repetions, and those whose inconsistency rates remain roughly constant. Good examples of the former are subjects 1, 11, 36, 39, 46 and 48. Subject 46 is a particularly interesting example - indeed this subject's inconsistency rate falls to zero and stays there. One might conclude from this that this subject had identified his or her preference functional towards the end of the experiment[4]. Subject 1 seems to be in the same category.

Some subjects' inconsistency rates stay roughly constant throughout the experiment. Good examples of this type of subject are subjects 6, 13, 14, 21, 26, 28, 42, 44 and 51. The behaviour of such subjects might be described either by a given preference functional with a given amount of error, or by a shifting preference functional. More light can be shed on these two alternatives in the analysis that follows. But first we introduce some important theoretical constructs.

4 Preference Functionals Investigated

The descriptive statistics contained in the section above are useful for getting a feel for the data, and, in particular, getting a feel for the magnitude of the variability in subjects' responses. We now need to relate our analysis to the theoretical literature on preference functionals (for decision-making under risk). We need to understand how well the various preference functionals help us to explain decision making in this context. We can then begin to explore the relationship between the variability that we have observed and to the correct modelling of the underlying preference functional. We begin by describing a (subset of) the preference functionals available in the literature. We then note that all of those considered here are deterministic preference functionals in the sense that, given a pair of prospects, the one that is preferred (using any of the preference functionals) is fixed. We then confront this with the results of our experiment - in which it is clear that stated preferences do change even with fixed pairwise choice problems. We discuss ways that this can be reconciled with the theoretical literature - without abandoning that literature. We are then in a position to use our data to analyse 'how well' the various preference functionals describe the behaviour in our experiment.

[4] Interestingly, this subject's behaviour at the end of the experiment can be precisely explained by a *lexicographic* (precisely minimax) preference function - by which we mean that the subject chooses that prospect for which the probability of the worst outcome is least.

The simplest preference functional for decision making under risk is Expected Value maximisation. This, of course, is a special case (Risk Neutrality: **RN**) of the more general Expected Utility (**EU**) maximisation. Both of these seem to have empirical shortcomings, particularly as judged from the perspective of experimental evidence. Indeed, these deficiencies have led to the development of alternative theoretical models, as discussed above. There are currently a large number of such alternatives, some more empirically valid than others. We did not feel it useful to explore all the proposed alternatives, but restricted attention to those that seem to have the highest empirical validity. We selected five other such functionals: Disppointment Aversion (**DA**); Prospective Reference (**PR**); Rank dependent with the Power weighting function (**RP**); Rank dependent with the Quiggin weighting function (**RQ**); and Weighted Utility (**WU**). We describe below the main features of the various preference functional: details of these can be found in (Hey, 1997). The two letter abbreviation is used to identify the functional. All the functionals are *holistic* and hence postulate a preference functional $V(.)$ which is used to rank prospects: hence $\mathbf{p}$ is preferred to $\mathbf{q}$ if and only if $V(\mathbf{p}) > V(\mathbf{q})$. The theories differ in their specification of $V(.)$[5] We shall limit the descriptions to the case relevant for the experiments described in this paper - where there are just 4 outcomes: x_1, x_2, x_3 and x_4, with respective probabilities p_1, p_2, p_3 and p_4.

RN : *Risk Neutrality* - subjects choose on the basis of expected value.

$$V(\mathbf{p}) = p_1 x_1 + p_2 x_2 + p_3 x_3 + p_4 x_4 \tag{1}$$

EU : *Expected Utility* - subjects choose on the basis of expected utility.

$$V(\mathbf{p}) = p_2 u(x_2) + p_3 u(x_3) + p_4 \tag{2}$$

DA : *Disappointment Aversion* - subjects choose on the basis of expected (modified) utility - where utility is modified *ex post* to take account of any disappointment or delight

[5] We normalise throughout with $u(x_1) = 0$ and $u(x_4) = 1$.

experienced.

$$V(\mathbf{p}) = \min(W_1, W_2, W_3) \tag{3}$$

where

$$W_1 = \frac{(1+\beta)p_2u(x_2) + (1+\beta)p_3u(x_3) + p_4}{1 + \beta p_1 + \beta p_2 + \beta p_3} \tag{4}$$

$$W_2 = \frac{(1+\beta)p_2u(x_2) + p_3u(x_3) + p_4}{1 + \beta p_1 + \beta p_2} \tag{5}$$

and

$$W_3 = \frac{p_2u(x_2) + p_3u(x_3) + p_4}{1 + \beta p_1} \tag{6}$$

PR : *Prospective Reference* - subjects choose on the basis of a weighted average of the expected utility calculated using the correct probabilities and the expected utility calculated using equal probabilities for all the non-null outcomes.

$$V(\mathbf{p}) = \lambda(p_2u(x_2) + p_3u(x_3) + p_4) + (1-\lambda)(a_2u(x_2) + a_3u(x_3) + a_4) \tag{7}$$

where $a_i = |a_i|/(a_i n(\mathbf{p}))$ and $n(\mathbf{p})$ is the number of non-zero elements in $\mathbf{p}$.

RP : *Rank dependent with Power weighting function* - subjects choose on the basis of expected utility where the (cumulative) probabilities are distorted by a weighting function which takes the power function form.

$$V(\mathbf{p}) = w(p_2 + p_3 + p_4)u(x_2) + w(p_3 + p_4)(u(x_3) - u(x_2)) + w(p_4)(1 - u(x_3)) \tag{8}$$

where $w(.)$ is the *power* function $w(p) = p^\gamma$.

RQ : *Rank dependent with Quiggin weighting function* - subjects choose on the basis of expected utility where the (cumulative) probabilities are distorted by a weighting function which takes the form advocated by Quiggin amongst others.

$$V(\mathbf{p}) = w(p_2 + p_3 + p_4)u(x_2) + w(p_3 + p_4)(u(x_3) - u(x_2)) + w(p_4)(1 - u(x_3)) \quad (9)$$

where $w(.)$ is the *'Quiggin'* function $w(p) = p^\gamma / [p^\gamma + (1-p)^\gamma]^{(1/\gamma)}$.

WU : *Weighted Utility* - subjects choose on the basis of expected weighted utility.

$$V(\mathbf{p}) = \frac{w_2 p_2 u(x_2) + w_3 p_3 u(x_3) + w_4 p_4}{w_1 p_1 + w_2 p_2 + w_3 p_3 + w_4 p_4} \quad (10)$$

All the generalisations of Expected Utility theory (**DA**, **PR**, **RP**, **RQ** and **WU**) involve one parameter extra to EU in the context of these experiments: **DA** has Gul's β parameter; **PR** has Viscusi's λ parameter; **RP** and **RQ** have the weighting function's γ parameter; and **WU** has the w weighting parameter.

5 Errors

The problem, as we have already identified, with these preference functionals, is that they are all deterministic. However it is clear from our experiment (and others) that subjects' behaviour is not deterministic, but that there is a lot of noise in the responses of subjects in such experiments. The clearest way to see this is to examine, as we have done above, the choices of subjects when confronted with the same pairwise choice question on different occasions (either within the same experimental setting or across different sessions): it is very clear that different responses are given on different occasions. How do we accommodate this? There are a number of ways.

One possibility is that subjects are responding to the experiment as a whole - considering it as a single response to 500 pairwise choice questions. We have discussed this above and have tried to argue that this story is unrealistic. In any case it would be extremely difficult to estimate, involving $2^{500} = 3.27 \times 10^{150}$ possibilities, an extraordinarily large number. To estimate the 'best fitting' preference functional under this scenario would require the computation of the likelihood (under some appropriate stochastic specification) at each of these $2^{500} = 3.27 \times 10^{150}$ points for each possible value of the parameter set. This is well beyond the capabilities of present day computors. (This emphasises the absurdity of this story - implying as it does that human beings can carry out such enormous calculations.)

There are simpler alternatives. The two obvious contenders are firstly that the subjects have *stochastic preference functions*, and secondly that subjects have deterministic preference functions but 'implement' them with error. The first of these has been explored elsewhere in a limited form[6] (see (Carbone, 1997)) but this is computationally difficult and involves fitting a large number of parameters. The second is computationally simpler and seems to be a simpler description of behaviour. We adopt that here, though some specification of the error structure must be selected. There are two possibilities that have already been explored in the literature: the White Noise specification of (Hey and Orme, 1994) and the Constant Probability specification of (Harless and Camerer, 1994). There are difficulties with both of these. The White Noise specification effectively assumes that subjects make measurement errors when calculating the value of their preference function, with the variance of the error constant across questions. This seems to be empirically invalid when the two prospects being compared are such that one prospect dominates the other[7]. The Constant Probability specification (which assumes that the probability of the subject mistakenly reporting his or her preference is constant across all questions) also suffers from this defect but additionally has the problem that it implies that the error probability is the same whether the prospects are far apart in the subject's preference or whether they are close together[8]. We adopt here the White Noise specification, though our results should be interpreted in the light of the possible deficiences of the specification. We assume that the White Noise error, the measurement error, is normally distributed with zero mean and constant variance. The variance, of course, is subject-specific and we estimate it along with the parameters of the preference functionals that we are fitting.

[6] Restricted to the Expected Utility functional.

[7] Experimental evidence seems to suggest that mistakes are very much less frequent with these types of questions than with others, where neither prospect dominates the other.

[8] Experimental evidence seems to suggest that the probability of making a mistake is less when the prospects are far apart.

6 Ways of Fitting the Data to the Set of Theories

Because of these ‘errors’, it is inevitable that in general, none of the preference functionals developed in the theoretical literature describes any of the subject’s behaviour precisely (for any parameters). We therefore need to find a method of choosing parameters for which particular preference functionals fit ‘as well as possible’. We use the method familiar to economists, that of choosing parameters which maximise the likelihood function - given the stochastic specification that we have adopted. Given that we assume that the White Noise error is normally distributed with zero mean and constant variance, we effectively have the Probit specification. However, since not all the preference functionals are linear in their parameters, we can not use standard probit estimation routines, because they typically assume linearity. Accordingly we have written our own Maximum Likelihood routines, using the GAUSS software package[9].

7 A Statistical Analysis of the Results

The descriptive analysis of Section 3 above gives us some insight into the behaviour of the subjects. However, if we want to answer the questions as to which of the various preference functionals best describes behaviour, and whether the repetition within the experiment leads to one particular best-fitting preference functional, we need to fit the various functionals to the observations. We proceed as outlined above: we fit the various preference functionals subject by subject, choosing the parameters of the respective functionals to maximise the (log of the) likelihood of observing the choices, given our White Noise assumption about the errors implicit in behaviour. We do this first repetition by repetition. For each subject, for each repetition, and for each preference functional we thus get a maximised log-likelihood which can be used to compare the goodness of fit of the respective preference functionals subject by subject and repetition by repetition. However, since the various preference functionals differ in terms of the number of parameters involved in their estimation, we need to correct the maximised log-likelihood for their varying degrees of

[9] Our estimation programs are available on request.

freedom. To be specific:

- Risk Neutrality involves one parameter (the standard deviation of the error) σ.
- Expected Utility involves the utility parameters u_2 (= the utility of £25) and u_3 (= the utility of £75)[10] and σ (the standard deviation of the error).
- Disappointment Aversion involves the utility parameters u_2 and u_3, σ and the 'disappointment aversion' parameter β.
- Prospective Reference involves the utility parameters u_2 and u_3, σ and Viscusi's parameter λ.
- Rank dependent with the Power weighting function involves the utility parameters u_2 and u_3, σ and the power exponent parameter γ.
- Rank dependent with the Quiggin weighting function involves the utility parameters u_2 and u_3, σ and Quiggin's parameter γ.
- Weighted Utility involves the utility parameters u_2 and u_3, σ and the weight parameter w.

So RN has one parameter, EU has 3, while all the others have 4. To correct the maximised log-likelihood for the numbers of parameters involved in the estimation (the number of degrees of freedom) we use the Akaike criterion, which involves penalising the maximised log-likelihood function by subtracting from it the number of degrees of freedom. Accordingly we obtain for each fitted preference functional a 'corrected' maximised log-likelihood, which can then be used to rank the various fitted preference functionals. Proceeding in this way we obtain Table 4. In this we list, for each subject and for each repetition, the preference functional for which the corrected maximised log-likelihood was highest - that is, the preference functional which came out best on what we term the Akaike Ranking. We also repeat the exercise for the preference functional fitted to all the data combined for each subject - that is the preference functional fitted to the data for all 500 observations

[10] Recall that we are normalising u_1 (= the utility of -£25) = 0 and u_4 (= the utility of £125) = 1.

for each subject. The preference functional which emerged as best on the Akaike Ranking is listed in the final column of Table 4.

Perhaps the only clear thing to emerge from Table 4 is the large amount of variability in the best-fitting[11] functional. There are some subjects for whom a stable preference functional seems to emerge by the end of the experiment - examples are subject 8 (WU), 18 (DA), 24 (RQ), 30 (RQ) and 38 (EU) - but for the majority, there is little, if any, convergence. It could, however, be argued that this test is too tough and hides small variations in the maximised log-likelihoods. Accordingly it might be useful to consider alternative analyses.

An alternative way to proceed is the following. We know that EU is nested within the more general preference functionals (DA, PR, RP, RQ and WU) - for each of these latter preference functionals one parameter restriction[12] reduces the more general functional to EU. Furthermore we know that RN is nested within EU - two parameter restrictions[13] reduce EU to RN. We might then first ask whether EU fits *significantly* better than RN (at some appropriate significance level), and then if it does, ask whether any of the more general preference functions fit *significantly* better than EU (at some appropriate significance level), and finally, if more than one does, choose the 'best' of these using the maximised log-likelihoods[14]. Proceeding in this way, we obtain Table 5 if we use a 5% significance level and Table 6 if we use a 1% significance level.

Table 5 shows two things: first, and rather obviously, EU emerges more often as the 'best-fitting' preference functional than in Table 4 but this is simply because the more general preference functionals are included in Table 5 only if they are *significantly* better that the less general. Secondly, but not invariably, we get some increased convergence. When we repeat the exercise using a 1% level of significance we get Table 6. Once again, and not surprisingly, we get an increased preponderance of cases in which EU emerges as the best-fitting functional. Rather more interestingly, we get an increased number of cases of convergence: subjects 1, 4, 8, 12, 18, 21, 25, 26, 31, 32, 34, 38, 45, 46, 48, 51 and 53 seem

[11] As defined by this Akaike criterion.

[12] Specifically, $\beta = 0$ for DA, $\lambda = 1$ for PR, $\gamma = 1$ for both RP and RQ and $w = 1$ for WU.

[13] Specifically, $u_2 = 1/3$ and $u_3 = 2/3$.

[14] Given that all the more general preference functionals have the same number of degrees of freedom.

all to converge to EU by the end of experiment. In contrast the number of subjects for whom a non-EU preference functional is the convergent state is reduced - though subject 5 converges to RQ, 27 to WU, 30 to RQ, 37 to RQ, 40 to RP and 50 to RQ.

The conclusion from these analyses is that there does appear to be some degree of convergence for some subjects, though for perhaps the majority the variability remains high.

The problem with these analyses is that they are essentially statistical in nature. We, as economists, might be more interested in the *economic* significance of the results. Given that the EU preference functional is much easier to apply to the economic analysis of behaviour, we might want to know how far wrong we might be if we use the EU functional rather than the alternatives in such applications. It is not obvious how we might answer this question as it depends upon the particular application. But we could ask how often we would make mistakes in the prediction of behaviour using the various preference functionals. This depends upon the predictions we are wanting to make. One possibility is to use the specific questions asked in this experiment - though it should be noted that the results of this analysis does depend upon the specific questions. It might be better to use some kind of generally-accepted set of questions - which can be used to test the various functionals - but such a set is not available and it is not clear how such a set could be constructed (and then made generally-acceptable).

The result of such an analysis are presented in Tables 7 through 16. There are two sets of tables here: Tables 7 through 11, which present the results of an analysis based on estimations repetition by repetition, and Tables 12 through 16 which present the results of an analysis based on estimations over all 5 repetitions. Let us begin with the first set.

For example, Table 7 refers to Repetition 1. For each subject, we count, for each preference functional, the number of questions (out of 100) for which (the best-fitting of) that preference functional incorrectly predicts the stated preference of that subject. For example, the RN preference functional predicts incorrectly 33 of the 100 responses of subject 1, th EU functional 7 of the 100, and so on. It is clear from Table 7 that Risk Neutrality does particularly badly at predicting behaviour but that the other functionals do much better. In general, the more general preference functionals make fewer mistakes in prediction, though this is hardly surprising[15].

Tables 7 through 11 present the analyses with estimations repetition by repetition. The column labelled 'maximum difference' calculates the difference between the number of incorrect predictions using the EU functional and the number of incorrect predictions using the functional which has the smallest number of such incorrect predictions. For some subjects this is 0 - implying that using EU instead of 'best-fitting' preference functional leads to no extra mistakes in prediction. However for some subjects this is as high as 13 (Subject 50 on Repetition 5), though it is below 10 in all except 2 cases. This suggests that using EU does not impose too great a cost in terms of predictability. One interesting question to ask is whether there is any relationship between this 'maximum difference' and the basic inconsistency in subjects' responses. Accordingly, we append to Tables 7 through 11 a column labelled 'average inconsistency' which is the average (one-quarter) of the 'total of these' column of Table 2. This indicates the average number of questions in which the subjects' responses changed between repetitions. It is apparent that there is no systematic relationship between this variable and the 'maximum difference' of the penultimate column. For example, on Repetition 1, (see Table 7 Subject 20, who has a very high variability in responses, has behaviour which is almost as well explained with the EU functional as with the more general functionals. In contrast, Subject 48, who has a relatively small variability, also has a relatively small 'maximum difference'.

A similar analysis is contained in Tables 12 through 16 though here the predictions are made (and the number of incorrect predictions calculated) using the appropriate best fittin preference functional *using all 500 observations*. Similarly the 'average inconsistency' column of this is the average (one-quarter) of the 'over all 5' column of Table 2. Recall this is the total number of questions, over all 5 repetitions, on which the subject gave different responses at some stage of the experiment. The final columns of Tables 12 through 16 can

[15] Though not inevitable: the functionals were fitted on the basis of maximising the likelihood, not on the basis of maximising the *score* - that is the number of correct predictions. Accordingly, it *could* be the case that the number of incorrect predictions is lower for EU, for example, than for one of the the more general functions (this, in fact happens for Subject 51 in Table 14). Moreover, it is not necessarily the case that the preference functional which is best in Table 4 has the smallest number of incorrect predictions - the reason, once again, being that the Akaike criterion maximises the corrected log-likelihood and not the score.

thus be considered as an average variability in responses over the experiment as a whole, and therefore is the appropriate 'conjugate' variable to go with the across-repetitions estimates of these tables. Here again there is no obvious relationship between the final two columns of Tables 12 through 16. There are examples of subjects where the 'maximum difference' is low, and where the 'average inconsistency' is high and also where it is low. For example, both Subjects 25 and 34 both have a 'maximum difference' of 0 but 'average inconsistencies' of 11.5 and 1.5 respectively. Indeed these tables indicate that there are different kinds of subjects with different kinds of behaviour.

Tables 12 through 16 are important in the sense that the numbers in the 'maximum difference' columns are typically very low. Occasionally they are negative (for reasons discussed in a footnote above) and they are all below 10. This indicates that if one is working with the combined data, to predict behaviour in individual contexts, then the use of the EU preference functional leads generally to errors which are remarkably low.

8 Conclusions

There are two important preliminary conclusions from this experiment: first, that there is a high degree of variability in subjects' responses, even in an experiment as simple as this; second, that there is a high degree of variability in subjects' behaviour during the experiment. We expand on these points below.

The variability of subjects' responses is high - the average percentage of differently answered questions between two repetitions is generally between 5 and 15 though it varies considerably across subjects. The average percentage of questions answered differently at some stage of the experiment is lower and is usually under 10, indicating that that the variability is limited to a subset of the questions. Across the repetitions the variability of responses declines for some subjects but stays constant for others (and indeed actually increases for a small number of subjects). For those subjects for whom the variability declines through time, it could be thought that these subjects are evolving their preference functional through the repetitions. There is some limited evidence that the majority of these are converging to the Expected Utility functional. For those subjects for whom the variability remains constant, it could be thought that their preferences are fixed (even though there is some noise in the expression of those preferences), but the evidence provided by the estimated preference functionals does not always confirm this[16].

In terms of the best-fitting preference functional, no clear picture emerges. If one confines attention to the final repetition, it would seem that EU is emerging as the best function (in Table 6 for 27 out of the 53 subjects EU is the 'best' functional), though there is a conflict between that and the 'best-fitting' functional using the combined data: if we compare the final two columns of Table 6 we see that there is often a conflict. Perhaps the appropriate strategy is to use the combined data when variablity is roughly constant and to use the data from the fifth repetition when the variability declines through the repetitions. Indeed there is clearly a problem for those subjects whose variability remains constant: the 'best-fitting' functional changes from repetition to repetition. A good example is Subject 6 for whom the 'best-fitting' functional (from Table 6) is EU, EU, RQ, RQ and WU for the 5 repetitions individually and WU overall. If WU is indeed this subject's true functional then the estimates from individual repetitions may be seriously misleading.

We seem to have the following preliminary conclusion: for those subjects whose variability is decreasing through the repetitions, we should take the estimated functional in the final repetition as their true functional; in contrast, for those subjects whose variability is roughly constant we should take the functional estimated over all 5 repetitions.

There is an additional conclusion: the increased errors in predictions using EU rather than the 'best-fitting' functional are generally low. Moreover, the magnitude of these errors *in comparison to the variability of the subjects' responses* are generally very small: if we compare the final two columns of Tables 7 through 16 we see that almost invariably the final column is larger than the penultimate. This means that the errors that the economist makes in predicting behaviour are generally of a smaller order of magnitude than the error that the subjects make themselves[17]. This clearly indicates that the way forward is to understand better the variability in subjects' responses: refining the deterministic preference functionals

[16] For example, Subject 42 has roughly constant variability, but the best-fitting functional varies: using the 1% criterion, DA, DA, RQ, WU and DA for the 5 repetitions and DA overall.

[17] Care should be taken in interpreting this conclusion in that the predictions are within-sample predictions, and generally the economist has to produce outside-sample predictions.

is not going to help if there is this innate randomness in subjects' behaviour. We need to understand better this innate randomness.

References

Carbone, E. (1997). Investigation of stochastic prefence theory using experimental data. *Economics Letters*, 57:305–311.

Cubitt, R., Starmer, C., and Sugden, R. (1998). On the validity of the random lottery incentive mechanism. *Experimental Economics*, 1:115–132.

Harless, D. and Camerer, C. (1994). The predictive utility of generalized expected utility theories. *Econometrica*, 62:1251–1290.

Hey, J. D. (1997). Experiments and the economics of individual decision making. In Kreps, D. M. and Wallis, K. F., editors, *Advances in Economics and Econometrics*, pages 171–205. Cambridge University Press.

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

Holt, C. A. (1986). Preference reversals and the independence axiom. *American Economic Review*, 76:508–515.

Karni, E. and Safra, Z. (1987). 'preference reversal' and the observability of preferences by experimental methods. *Econometrica*, 55:675–685.

Machina, M. (1982). ‘expected utility’ analysis without the independence axiom. *Econometrica*, 50:277–323.

List of Tables

Table 1: The 100 Pairwise Choice Questions

Question Number	Choice 1				Choice 2			
	p_1	p_2	p_3	p_4	p_1	p_2	p_3	p_4
1	.0	.0	.875	.125	.0	.125	.0	.875
2	.0	.0	.875	.125	.0	.125	.0	.875
3	.0	.0	.875	.125	.0	.125	.5	.375
4	.0	.0	.875	.125	.0	.375	.0	.625
5	.0	.0	.875	.125	.0	.375	.125	.5
6	.0	.0	.875	.125	.0	.375	.25	.375
7	.0	.0	.875	.125	.0	.625	.0	.375
8	.0	.125	.5	.375	.0	.375	.0	.625
9	.0	.125	.5	.375	.0	.375	.125	.5
10	.0	.125	.875	.0	.0	.375	.0	.625
11	.0	.125	.875	.0	.0	.375	.125	.5
12	.0	.125	.875	.0	.0	.375	.25	.375
13	.0	.125	.875	.0	.0	.375	.5	.125
14	.0	.125	.875	.0	.0	.625	.0	.375
15	.0	.125	.875	.0	.0	.875	.0	.125
16	.0	.25	.75	.0	.0	.375	.0	.625
17	.0	.25	.75	.0	.0	.375	.125	.5
18	.0	.25	.75	.0	.0	.375	.25	.375
19	.0	.25	.75	.0	.0	.375	.5	.125
20	.0	.25	.75	.0	.0	.375	.5	.125
21	.0	.25	.75	.0	.0	.625	.0	.375
22	.0	.25	.75	.0	.0	.875	.0	.125
23	.0	.375	.5	.125	.0	.625	.0	.375
24	.0	.125	.875	.0	.0	.25	.75	.0
25	.0	.375	.125	.5	.0	.375	.25	.375
26	.0	.0	.5	.5	.125	.0	.25	.625
27	.0	.0	.5	.5	.125	.0	.25	.625
28	.0	.0	.875	.125	.125	.0	.25	.625
29	.0	.0	.875	.125	.125	.0	.625	.25
30	.0	.0	.875	.125	.375	.0	.375	.25
31	.0	.0	.875	.125	.5	.0	.0	.5
32	.0	.0	.875	.125	.75	.0	.0	.25
33	.0	.0	1.0	.0	.125	.0	.25	.625
34	.0	.0	1.0	.0	.125	.0	.625	.25
35	.0	.0	1.0	.0	.375	.0	.375	.25
36	.0	.0	1.0	.0	.5	.0	.0	.5
37	.0	.0	1.0	.0	.75	.0	.0	.25
38	.0	.0	1.0	.0	.75	.0	.0	.25
39	.0	.0	1.0	.0	.75	.0	.125	.125
40	.125	.0	.625	.25	.5	.0	.0	.5
41	.25	.0	.75	.0	.375	.0	.375	.25
42	.25	.0	.75	.0	.5	.0	.0	.5
43	.25	.0	.75	.0	.75	.0	.0	.25
44	.25	.0	.75	.0	.75	.0	.125	.125
45	.375	.0	.375	.25	.5	.0	.0	.5
46	.375	.0	.625	.0	.5	.0	.0	.5
47	.375	.0	.625	.0	.75	.0	.0	.25
48	.375	.0	.625	.0	.75	.0	.125	.125
49	.25	.0	.75	.0	.375	.0	.625	.0
50	.75	.0	.0	.25	.75	.0	.125	.125
51	.0	.75	.0	.25	.25	.375	.0	.375
52	.0	.75	.0	.25	.375	.125	.0	.5
53	.0	.75	.0	.25	.625	.0	.0	.375
54	.0	.875	.0	.125	.25	.375	.0	.375

Table 1: The 100 Pairwise Choice Questions

Question	Choice 1				Choice 2			
Number	p_1	p_2	p_3	p_4	p_1	p_2	p_3	p_4
55	.0	.875	.0	.125	.375	.125	.0	.5
56	.0	.875	.0	.125	.5	.25	.0	.25
57	.0	.875	.0	.125	.625	.0	.0	.375
58	.0	.875	.0	.125	.625	.125	.0	.25
59	.125	.75	.0	.125	.25	.375	.0	.375
60	.125	.75	.0	.125	.375	.125	.0	.5
61	.125	.75	.0	.125	.5	.25	.0	.25
62	.125	.75	.0	.125	.625	.0	.0	.375
63	.125	.75	.0	.125	.625	.125	.0	.25
64	.125	.875	.0	.0	.25	.375	.0	.375
65	.125	.875	.0	.0	.375	.125	.0	.5
66	.125	.875	.0	.0	.5	.25	.0	.25
67	.125	.875	.0	.0	.625	.0	.0	.375
68	.125	.875	.0	.0	.625	.125	.0	.25
69	.125	.875	.0	.0	.75	.125	.0	.125
70	.125	.875	.0	.0	.875	.0	.0	.125
71	.125	.875	.0	.0	.875	.0	.0	.125
72	.25	.375	.0	.375	.375	.125	.0	.5
73	.5	.25	.0	.25	.625	.0	.0	.375
74	.5	.25	.0	.25	.625	.0	.0	.375
75	.0	.75	.0	.25	.125	.75	.0	.125
76	.0	.75	.25	.0	.125	.0	.875	.0
77	.0	.75	.25	.0	.125	.375	.5	.0
78	.0	.75	.25	.0	.375	.125	.5	.0
79	.0	.75	.25	.0	.375	.25	.375	.0
80	.0	.75	.25	.0	.5	.0	.5	.0
81	.0	.75	.25	.0	.5	.125	.375	.0
82	.0	1.0	.0	.0	.125	.0	.875	.0
83	.0	1.0	.0	.0	.125	.375	.5	.0
84	.0	1.0	.0	.0	.25	.625	.125	.0
85	.0	1.0	.0	.0	.375	.125	.5	.0
86	.0	1.0	.0	.0	.375	.25	.375	.0
87	.0	1.0	.0	.0	.5	.0	.5	.0
88	.0	1.0	.0	.0	.5	.0	.5	.0
89	.0	1.0	.0	.0	.5	.125	.375	.0
90	.0	1.0	.0	.0	.75	.125	.125	.0
91	.25	.625	.125	.0	.375	.125	.5	.0
92	.25	.625	.125	.0	.375	.25	.375	.0
93	.25	.625	.125	.0	.5	.0	.5	.0
94	.25	.625	.125	.0	.5	.125	.375	.0
95	.375	.25	.375	.0	.5	.0	.5	.0
96	.375	.25	.375	.0	.5	.0	.5	.0
97	.375	.625	.0	.0	.5	.0	.5	.0
98	.375	.625	.0	.0	.5	.125	.375	.0
99	.375	.625	.0	.0	.75	.125	.125	.0
100	.375	.125	.5	.0	.5	.125	.375	.0

Table 2: Inconsistency rates between repetitions

Subject	Repetition				total	over
Number	1 to 2	2 to 3	3 to 4	4 to 5	of these	all 5
1	9	3	4	1	17	9
2	13	9	5	7	34	20
3	8	11	9	8	36	19
4	8	10	9	5	32	18
5	10	14	8	4	36	18
6	13	13	11	14	51	29
7	16	9	9	7	41	25
8	12	16	17	14	59	32
9	17	12	10	10	49	24
10	18	10	4	6	38	28
11	9	5	3	2	19	12
12	6	4	3	2	15	8
13	13	15	13	12	53	26
14	11	13	9	13	46	27
15	12	12	10	11	45	28
16	10	7	13	12	42	23
17	12	12	14	23	61	38
18	12	10	10	8	40	23
19	9	12	17	15	53	32
20	29	22	18	22	91	48
21	15	14	10	14	53	30
22	5	8	11	7	31	17
23	7	8	11	16	42	25
24	13	12	20	11	56	30
25	22	19	25	17	83	46
26	14	17	19	13	63	40
27	22	16	19	10	67	39
28	3	7	11	7	28	16
29	9	12	8	5	34	20
30	7	13	11	5	36	19
31	13	14	7	9	43	22
32	6	6	3	4	19	10
33	8	9	8	6	31	18
34	4	3	3	3	13	6
35	4	9	13	13	39	23
36	11	6	3	3	23	18
37	12	11	5	10	38	22
38	22	16	12	11	61	31
39	8	12	3	2	25	18
40	12	6	6	8	32	17
41	15	12	17	14	58	30
42	13	15	16	12	56	31
43	11	8	15	14	48	26
44	4	3	4	4	15	9
45	7	1	2	3	13	11
46	3	1	0	0	4	3
47	9	9	5	5	28	15
48	14	9	4	4	31	22
49	10	14	10	6	40	23
50	9	6	9	9	33	18
51	14	9	12	10	45	22
52	19	18	14	14	65	36
53	14	9	14	9	46	25

Table 3: Violations of Dominance

Question Number	Repetition				
	1	2	3	4	5
24	2	1	0	2	1
25	1	5	3	2	3
49	1	0	0	0	0
50	0	0	0	2	0
75	0	0	0	0	1
100	0	0	0	0	0

Table 4: Best models on Akaike Ranking

Subject	Repetition					all
Number	1	2	3	4	5	combined
1	RQ	PR	PR	RQ	EU	PR
2	PR	RQ	RQ	RQ	RQ	PR
3	RQ	DA	WU	RP	WU	RQ
4	WU	EU	RQ	EU	WU	WU
5	WU	RQ	DA	RQ	RQ	RQ
6	WU	RQ	RQ	RQ	WU	WU
7	DA	WU	EU	WU	RP	DA
8	WU	RQ	WU	WU	WU	WU
9	RQ	WU	EU	RP	RP	WU
10	EU	RP	WU	WU	RP	RP
11	RP	DA	RQ	RQ	RQ	RP
12	RQ	RP	DA	EU	DA	RQ
13	RQ	RP	PR	PR	PR	PR
14	PR	WU	WU	WU	DA	WU
15	EU	RQ	DA	DA	WU	WU
16	WU	RQ	DA	PR	RQ	RQ
17	WU	WU	RP	WU	PR	WU
18	WU	RP	DA	DA	DA	DA
19	DA	RQ	RQ	WU	PR	RQ
20	RQ	EU	RP	WU	RP	DA
21	EU	RP	PR	PR	DA	PR
22	DA	RP	RP	RP	RP	RP
23	DA	RQ	DA	RQ	PR	RQ
24	RQ	RQ	RQ	RQ	RQ	RQ
25	PR	WU	WU	PR	WU	WU
26	WU	RP	RP	EU	EU	RP
27	RP	WU	WU	WU	WU	WU
28	RQ	RQ	DA	PR	WU	RQ
29	RP	RP	PR	RQ	RQ	PR
30	WU	RQ	RQ	RQ	RQ	RQ
31	WU	EU	EU	EU	WU	WU
32	EU	DA	RP	RP	RP	RP
33	RQ	WU	RQ	DA	RQ	RQ
34	EU	RP	EU	WU	EU	RP
35	RQ	RQ	DA	WU	EU	RQ
36	RQ	RP	PR	RQ	RP	DA
37	PR	RQ	RQ	RQ	RQ	PR
38	EU	DA	EU	EU	EU	EU
39	RP	EU	RQ	RQ	PR	PR
40	PR	PR	PR	RP	RP	PR
41	WU	PR	WU	PR	RQ	PR
42	DA	DA	RQ	WU	DA	DA
43	EU	EU	EU	RP	EU	EU
44	RQ	RP	RQ	RP	RQ	RP
45	PR	DA	DA	RQ	DA	WU
46	RQ	EU	EU	EU	EU	RQ
47	DA	RQ	DA	DA	RQ	RQ
48	EU	DA	DA	WU	DA	WU
49	PR	RP	EU	RQ	PR	EU
50	RQ	RP	RP	RQ	RQ	RQ
51	EU	EU	RP	EU	RP	EU
52	EU	PR	PR	PR	DA	PR
53	DA	WU	PR	RP	RP	RP

Table 5: Best models on significance (5 per cent) and ranking

Subject	Repetition					all
Number	1	2	3	4	5	combined
1	RQ	EU	PR	EU	EU	PR
2	PR	RQ	RQ	RQ	RQ	PR
3	EU	DA	WU	RP	WU	RQ
4	EU	EU	EU	EU	WU	WU
5	WU	RQ	DA	RQ	RQ	RQ
6	WU	EU	RQ	RQ	WU	WU
7	EU	WU	EU	WU	RP	EU
8	WU	RQ	WU	WU	EU	WU
9	RQ	WU	EU	RP	RP	WU
10	EU	RP	WU	WU	RP	RP
11	RP	DA	RQ	RQ	RQ	RP
12	RQ	RP	DA	EU	EU	RQ
13	RQ	RP	PR	PR	PR	PR
14	PR	WU	WU	WU	DA	WU
15	EU	RQ	DA	DA	WU	WU
16	WU	RQ	DA	PR	RQ	RQ
17	WU	WU	RP	WU	PR	WU
18	WU	EU	EU	EU	EU	DA
19	DA	RQ	RQ	WU	EU	RQ
20	RQ	EU	RP	WU	RP	DA
21	EU	EU	EU	EU	DA	EU
22	DA	RP	RP	RP	RP	RP
23	DA	EU	EU	RQ	EU	RQ
24	RQ	RQ	RQ	RQ	RQ	RQ
25	PR	EU	WU	EU	WU	WU
26	EU	RP	RP	EU	EU	RP
27	RP	WU	WU	WU	WU	WU
28	RQ	RQ	EU	PR	WU	RQ
29	EU	RP	PR	RQ	RQ	PR
30	WU	RQ	RQ	RQ	RQ	RQ
31	WU	EU	EU	EU	WU	WU
32	EU	DA	RP	RP	EU	RP
33	RQ	WU	RQ	EU	RQ	RQ
34	EU	RP	EU	EU	EU	RP
35	RQ	RQ	DA	WU	EU	RQ
36	RQ	EU	EU	RQ	RP	DA
37	PR	RQ	RQ	RQ	RQ	PR
38	EU	EU	EU	EU	EU	EU
39	RP	EU	RQ	RQ	PR	PR
40	PR	PR	PR	RP	RP	PR
41	WU	EU	WU	PR	RQ	PR
42	DA	DA	RQ	WU	DA	DA
43	EU	EU	EU	RP	EU	EU
44	RQ	RP	RQ	RP	RQ	RP
45	PR	DA	DA	EU	EU	WU
46	RQ	EU	EU	EU	EU	RQ
47	DA	RQ	DA	DA	RQ	RQ
48	EU	DA	EU	WU	DA	EU
49	PR	EU	EU	EU	PR	EU
50	RQ	RP	RP	RQ	RQ	RQ
51	EU	EU	RP	EU	EU	EU
52	EU	PR	PR	PR	DA	PR
53	DA	WU	PR	RP	RP	EU

Table 6: Best models on significance (1 per cent) and Ranking

Subject	Repetition					all
Number	1	2	3	4	5	combined
1	RQ	EU	EU	EU	EU	PR
2	PR	RQ	RQ	RQ	RQ	PR
3	EU	EU	WU	RP	EU	RQ
4	EU	EU	EU	EU	EU	WU
5	EU	EU	DA	RQ	RQ	RQ
6	EU	EU	RQ	RQ	WU	WU
7	EU	EU	EU	WU	EU	EU
8	WU	RQ	EU	EU	EU	WU
9	RQ	EU	EU	RP	EU	WU
10	EU	EU	WU	WU	RP	RP
11	RP	DA	RQ	RQ	RQ	RP
12	RQ	EU	DA	EU	EU	RQ
13	EU	RP	PR	PR	PR	PR
14	EU	WU	WU	EU	DA	WU
15	EU	EU	DA	DA	WU	WU
16	WU	RQ	DA	PR	RQ	RQ
17	WU	EU	RP	WU	PR	WU
18	WU	EU	EU	EU	EU	DA
19	DA	RQ	RQ	WU	EU	RQ
20	RQ	EU	RP	WU	EU	DA
21	EU	EU	EU	EU	EU	EU
22	DA	EU	RP	RP	EU	RP
23	DA	EU	EU	RQ	EU	RQ
24	RQ	RQ	EU	RQ	RQ	RQ
25	EU	EU	EU	EU	EU	WU
26	EU	RP	EU	EU	EU	RP
27	RP	WU	WU	WU	WU	WU
28	RQ	RQ	EU	PR	WU	RQ
29	EU	EU	PR	RQ	RQ	PR
30	WU	RQ	EU	RQ	RQ	RQ
31	EU	EU	EU	EU	EU	WU
32	EU	EU	EU	EU	EU	RP
33	RQ	WU	RQ	EU	RQ	RQ
34	EU	EU	EU	EU	EU	EU
35	RQ	RQ	EU	WU	EU	RQ
36	RQ	EU	EU	RQ	RP	DA
37	PR	RQ	RQ	RQ	RQ	PR
38	EU	EU	EU	EU	EU	EU
39	RP	EU	RQ	RQ	PR	PR
40	PR	EU	PR	RP	RP	PR
41	WU	EU	WU	PR	RQ	PR
42	DA	DA	RQ	WU	DA	DA
43	EU	EU	EU	RP	EU	EU
44	RQ	RP	RQ	RP	RQ	RP
45	EU	DA	DA	EU	EU	WU
46	RQ	EU	EU	EU	EU	RQ
47	DA	RQ	DA	DA	RQ	RQ
48	EU	DA	EU	EU	EU	EU
49	EU	EU	EU	EU	PR	EU
50	RQ	RP	RP	RQ	RQ	RQ
51	EU	EU	RP	EU	EU	EU
52	EU	PR	EU	PR	EU	PR
53	EU	EU	EU	EU	EU	EU

Table 7: Incorrect Predictions, Individual Estimates, Repetition 1

Subject	Repetition							maximum	average
Number	RN	EU	DA	PR	RP	RQ	WU	difference	inconsistency
1	33	7	7	6	2	3	7	5	4.25
2	29	13	14	9	9	10	8	5	8.50
3	34	7	6	7	7	7	5	2	9.00
4	21	16	16	15	16	15	12	4	8.00
5	31	5	5	5	5	3	3	2	9.00
6	18	12	12	8	12	10	11	4	12.75
7	33	8	10	9	11	9	7	1	10.25
8	25	9	9	7	8	5	7	4	14.75
9	28	14	13	12	16	13	14	2	12.25
10	35	7	7	6	6	5	7	2	9.50
11	33	9	9	8	6	9	9	3	4.75
12	35	5	5	5	2	1	3	4	3.75
13	15	8	8	8	8	7	8	1	13.25
14	19	7	7	10	8	10	10	0	11.50
15	9	8	8	8	7	6	7	2	11.25
16	26	9	6	7	9	8	6	3	10.50
17	15	10	10	9	9	10	3	7	15.25
18	16	12	9	11	12	8	7	5	10.00
19	30	16	12	12	18	10	14	6	13.25
20	29	22	21	23	23	22	22	1	22.75
21	21	15	10	11	10	12	14	5	13.25
22	17	6	4	7	5	6	7	2	7.75
23	22	13	9	11	13	11	10	4	10.50
24	29	12	9	8	12	8	12	4	14.00
25	32	15	14	15	15	15	14	1	20.75
26	11	9	9	8	9	7	9	2	15.75
27	22	15	15	11	13	15	15	4	16.75
28	31	9	6	4	5	2	6	7	7.00
29	26	7	7	7	8	5	6	2	8.50
30	24	12	11	11	10	10	10	2	9.00
31	31	12	10	11	11	12	7	5	10.75
32	34	5	5	5	5	5	5	0	4.75
33	32	12	14	9	10	3	7	9	7.75
34	6	2	2	2	4	2	2	0	3.25
35	36	3	3	3	2	2	3	1	9.75
36	30	3	6	4	3	6	3	0	5.75
37	32	12	11	9	11	8	11	4	9.50
38	33	12	12	12	12	12	14	0	15.25
39	21	9	9	8	10	9	10	1	6.25
40	24	13	13	6	12	5	11	8	8.00
41	13	19	19	13	15	17	12	7	14.50
42	34	19	14	18	15	15	26	5	14.00
43	21	8	8	7	8	7	7	1	12.00
44	37	7	7	7	4	3	5	4	3.75
45	34	8	8	6	8	7	7	2	3.25
46	38	4	4	3	2	0	0	4	1.00
47	27	12	10	11	13	10	7	5	7.00
48	20	11	12	10	11	11	11	1	7.75
49	23	13	12	8	13	9	11	5	10.00
50	40	11	9	11	8	6	11	5	8.25
51	30	9	9	9	9	8	7	2	11.25
52	18	16	16	16	16	16	16	0	16.25
53	31	9	11	11	10	10	11	-1	11.50

Table 8: Incorrect Predictions, Individual Estimates, Repetition 2

Subject	Repetition							maximum	average
Number	RN	EU	DA	PR	RP	RQ	WU	difference	inconsistency
1	38	3	3	2	3	2	3	1	4.25
2	33	10	9	6	8	6	10	4	8.50
3	31	8	7	6	10	7	8	2	9.00
4	20	15	15	15	15	15	14	1	8.00
5	33	9	8	8	9	8	7	2	9.00
6	17	14	14	15	14	15	18	0	12.75
7	38	8	8	9	9	9	7	1	10.25
8	25	16	16	13	11	14	15	5	14.75
9	27	9	15	10	11	11	8	1	12.25
10	25	13	13	8	12	10	13	5	9.50
11	32	5	1	5	4	2	1	4	4.75
12	35	2	2	2	0	2	2	2	3.75
13	18	14	14	15	12	15	14	2	13.25
14	21	11	9	9	9	9	5	6	11.50
15	17	11	12	9	11	9	7	4	11.25
16	29	7	6	4	7	3	5	4	10.50
17	21	13	13	14	13	15	9	4	15.25
18	22	12	12	12	8	11	10	4	10.00
19	27	17	15	12	16	12	14	5	13.25
20	30	19	20	19	17	19	19	2	22.75
21	15	10	10	8	11	10	13	2	13.25
22	18	11	8	10	9	10	9	3	7.75
23	19	5	6	7	8	3	4	2	10.50
24	26	9	7	7	9	7	5	4	14.00
25	27	16	17	16	20	18	14	2	20.75
26	14	12	14	13	10	13	15	2	15.75
27	13	14	14	13	17	13	15	1	16.75
28	34	5	1	3	2	0	1	5	7.00
29	27	8	8	6	8	5	7	3	8.50
30	27	14	9	7	14	5	7	9	9.00
31	26	11	10	10	11	10	12	1	10.75
32	33	5	3	4	4	3	4	2	4.75
33	28	9	8	7	7	8	6	3	7.75
34	7	2	3	1	4	2	3	1	3.25
35	35	4	2	4	3	2	4	2	9.75
36	31	2	1	2	1	2	1	1	5.75
37	41	17	15	8	15	10	14	9	9.50
38	34	20	18	18	18	19	16	4	15.25
39	20	10	10	9	10	10	11	1	6.25
40	23	9	7	11	10	11	9	2	8.00
41	17	9	9	9	9	8	9	1	14.50
42	30	11	8	12	14	8	8	3	14.00
43	17	10	10	10	10	9	9	1	12.00
44	35	4	4	3	1	4	4	3	3.75
45	37	5	0	4	3	0	0	5	3.25
46	41	1	1	1	1	1	1	0	1.00
47	26	18	12	12	17	12	11	7	7.00
48	30	3	1	4	3	2	2	2	7.75
49	25	9	9	9	10	11	9	0	10.00
50	38	6	6	6	5	4	6	2	8.25
51	31	7	7	7	7	7	7	0	11.25
52	19	16	16	12	13	13	14	4	16.25
53	30	5	5	5	5	5	7	0	11.50

Table 9: Incorrect Predictions, Individual Estimates, Repetition 3

Subject Number	Repetition							maximum difference	average inconsistency
	RN	EU	DA	PR	RP	RQ	WU		
1	37	3	3	1	3	2	3	2	4.25
2	33	10	7	5	7	6	10	5	8.50
3	31	6	6	7	5	5	4	2	9.00
4	18	12	12	11	13	10	7	5	8.00
5	33	10	8	9	8	7	9	3	9.00
6	15	13	14	11	11	11	14	2	12.75
7	40	6	6	5	6	6	6	1	10.25
8	20	19	19	18	18	16	15	4	14.75
9	30	10	10	10	11	11	11	0	12.25
10	25	6	6	6	7	6	5	1	9.50
11	36	7	7	6	2	0	2	7	4.75
12	35	5	3	5	5	2	4	3	3.75
13	12	8	9	4	7	4	6	4	13.25
14	27	12	12	8	13	10	8	4	11.50
15	20	11	10	13	13	11	5	6	11.25
16	30	6	3	5	8	4	3	3	10.50
17	19	13	13	13	7	12	6	7	15.25
18	24	8	6	8	6	8	8	2	10.00
19	25	15	14	10	13	9	6	9	13.25
20	34	14	9	12	9	12	21	5	22.75
21	16	11	11	14	13	12	12	0	13.25
22	20	11	12	11	10	12	13	1	7.75
23	20	13	8	9	13	10	9	5	10.50
24	30	9	7	9	9	7	10	2	14.00
25	23	16	15	19	19	17	15	1	20.75
26	13	15	16	15	8	15	17	7	15.75
27	16	16	16	18	15	18	14	2	16.75
28	36	13	10	13	11	11	12	3	7.00
29	32	7	6	6	8	6	6	1	8.50
30	26	13	13	10	12	9	11	4	9.00
31	38	0	0	0	0	0	0	0	10.75
32	35	5	5	5	2	3	5	3	4.75
33	31	8	6	6	6	4	4	4	7.75
34	6	1	2	1	1	1	2	0	3.25
35	33	5	2	5	5	3	2	3	9.75
36	32	4	4	2	2	2	4	2	5.75
37	38	8	3	7	8	4	3	5	9.50
38	28	7	6	7	7	7	8	1	15.25
39	20	8	8	5	11	6	7	3	6.25
40	24	10	10	6	5	6	9	5	8.00
41	15	13	12	13	13	12	12	1	14.50
42	19	11	11	8	9	8	9	3	14.00
43	22	6	6	8	5	7	6	1	12.00
44	35	6	2	4	1	1	2	5	3.75
45	38	4	0	3	4	0	0	4	3.25
46	40	0	0	0	0	0	0	0	1.00
47	26	18	18	15	22	13	10	8	7.00
48	35	7	6	7	7	6	8	1	7.75
49	30	7	7	6	7	6	7	1	10.00
50	34	3	3	4	3	3	5	0	8.25
51	27	6	6	3	2	3	5	4	11.25
52	22	13	14	11	13	12	13	2	16.25
53	31	8	8	5	7	7	6	3	11.50

Table 10: Incorrect Predictions, Individual Estimates, Repetition 4

Subject	Repetition							maximum	average
Number	RN	EU	DA	PR	RP	RQ	WU	difference	inconsistency
1	37	3	2	2	2	2	3	1	4.25
2	36	10	11	7	8	6	11	4	8.50
3	36	6	6	6	5	5	6	1	9.00
4	22	13	13	12	10	11	12	3	8.00
5	36	7	4	6	5	4	4	3	9.00
6	16	14	14	8	12	9	13	6	12.75
7	37	5	4	5	5	5	4	1	10.25
8	22	15	15	12	15	14	12	3	14.75
9	28	10	8	9	10	9	9	2	12.25
10	27	4	4	5	4	6	5	0	9.50
11	36	7	7	4	5	2	3	5	4.75
12	34	0	0	0	0	0	0	0	3.75
13	14	14	14	9	7	10	9	7	13.25
14	24	12	13	12	11	13	12	1	11.50
15	22	7	8	8	9	7	8	0	11.25
16	25	11	9	7	9	6	9	5	10.50
17	21	12	12	11	13	9	11	3	15.25
18	22	10	9	9	8	9	9	2	10.00
19	15	11	11	6	11	7	4	7	13.25
20	34	13	10	15	11	12	17	3	22.75
21	16	13	13	15	13	15	15	0	13.25
22	12	9	9	6	6	9	8	3	7.75
23	22	13	12	10	13	10	10	3	10.50
24	29	23	16	14	22	17	17	9	14.00
25	23	19	19	19	20	18	20	1	20.75
26	25	5	5	5	5	5	4	1	15.75
27	17	17	17	16	16	16	12	5	16.75
28	35	5	5	2	3	3	5	3	7.00
29	35	7	7	6	6	6	5	2	8.50
30	33	10	8	5	6	5	7	5	9.00
31	37	8	7	8	7	8	8	1	10.75
32	34	4	4	3	3	4	5	1	4.75
33	33	7	8	7	9	8	5	2	7.75
34	7	2	2	1	1	1	1	1	3.25
35	30	9	9	9	9	10	6	3	9.75
36	34	2	2	1	2	0	2	2	5.75
37	39	9	8	4	8	6	7	5	9.50
38	30	10	9	10	11	9	10	1	15.25
39	19	5	5	3	7	4	3	2	6.25
40	28	13	13	10	9	11	10	4	8.00
41	16	16	16	14	16	14	16	2	14.50
42	26	11	8	16	14	10	4	7	14.00
43	25	11	9	10	12	11	11	2	12.00
44	36	6	6	5	3	2	6	4	3.75
45	40	2	2	3	2	3	2	0	3.25
46	40	0	0	0	0	0	0	0	1.00
47	26	20	16	15	19	15	10	10	7.00
48	33	5	4	4	3	4	5	2	7.75
49	29	6	6	5	4	5	4	2	10.00
50	34	13	10	9	17	8	14	5	8.25
51	23	8	8	7	9	8	8	1	11.25
52	20	15	12	10	12	8	10	7	16.25
53	33	14	14	12	10	13	9	5	11.50

Table 11: Incorrect Predictions, Individual Estimates, Repetition 5

Subject	Repetition							maximum	average
Number	RN	EU	DA	PR	RP	RQ	WU	difference	inconsistency
1	38	0	0	0	0	0	0	0	4.25
2	35	14	15	12	13	10	11	4	8.50
3	33	7	5	4	7	3	5	4	9.00
4	21	14	14	16	12	14	10	4	8.00
5	33	3	3	3	2	4	2	1	9.00
6	11	13	14	15	14	13	12	1	12.75
7	36	6	6	5	5	6	6	1	10.25
8	17	15	15	16	15	15	11	4	14.75
9	32	3	3	3	4	5	3	0	12.25
10	25	9	9	9	8	8	10	1	9.50
11	36	7	7	4	5	0	20	7	4.75
12	36	1	1	1	1	1	1	0	3.75
13	17	11	10	4	9	4	5	7	13.25
14	28	16	13	13	16	13	16	3	11.50
15	14	8	8	8	8	9	3	5	11.25
16	27	16	11	8	16	8	41	8	10.50
17	24	27	27	24	25	27	24	3	15.25
18	26	5	4	6	5	5	5	1	10.00
19	30	11	10	11	10	10	10	1	13.25
20	31	14	15	16	16	18	18	-1	22.75
21	17	7	6	7	6	7	7	1	13.25
22	16	10	8	9	8	7	7	3	7.75
23	18	7	7	5	6	6	8	2	10.50
24	27	15	14	13	16	16	11	4	14.00
25	30	7	7	6	7	7	7	1	20.75
26	31	11	11	11	8	10	10	3	15.75
27	10	18	18	12	18	15	11	7	16.75
28	31	7	8	8	6	5	3	4	7.00
29	33	6	6	5	5	3	4	3	8.50
30	30	11	9	8	11	7	10	4	9.00
31	32	4	3	4	4	4	4	1	10.75
32	35	6	6	6	3	4	6	3	4.75
33	29	11	8	6	11	7	7	5	7.75
34	6	1	2	1	1	1	2	0	3.25
35	31	11	10	8	10	9	7	4	9.75
36	35	2	2	4	2	2	3	0	5.75
37	40	13	7	10	13	5	11	8	9.50
38	31	4	4	4	4	4	4	0	15.25
39	21	5	7	4	7	6	5	1	6.25
40	26	8	8	5	6	5	9	3	8.00
41	17	6	7	6	5	5	3	3	14.50
42	27	10	6	10	9	9	6	4	14.00
43	29	8	8	7	8	7	6	2	12.00
44	37	7	3	8	4	2	3	5	3.75
45	41	3	3	2	3	3	3	1	3.25
46	40	0	0	0	0	0	0	0	1.00
47	24	16	15	12	20	13	11	5	7.00
48	35	5	3	6	7	5	7	2	7.75
49	27	3	3	0	3	0	3	3	10.00
50	39	14	12	9	14	1	18	13	8.25
51	27	7	7	5	5	6	6	2	11.25
52	16	14	13	11	15	10	13	4	16.25
53	25	8	8	8	4	8	6	4	11.50

Table 12: Incorrect Predictions, Combined Estimates, Repetition 1

Subject	Repetition							maximum	average
Number	RN	EU	DA	PR	RP	RQ	WU	difference	inconsistency
1	30	7	7	7	7	6	7	1	2.25
2	21	15	15	10	11	9	15	6	5.00
3	29	10	9	9	9	8	9	2	4.75
4	13	16	16	16	16	15	12	4	4.50
5	24	7	9	9	9	5	7	2	4.50
6	11	11	11	11	13	15	17	0	7.25
7	26	17	17	18	17	17	17	0	6.25
8	19	8	9	9	7	8	10	1	8.00
9	23	14	13	13	14	12	11	3	6.00
10	29	12	12	12	14	12	12	0	7.00
11	27	10	10	12	7	11	10	3	3.00
12	28	5	5	5	5	5	5	0	2.00
13	6	12	12	7	9	6	7	6	6.50
14	14	9	12	10	9	11	10	0	6.75
15	3	15	16	16	15	15	15	0	7.00
16	19	10	11	11	10	11	8	2	5.75
17	7	15	15	13	17	14	14	2	9.50
18	10	10	10	12	12	11	10	0	5.75
19	23	15	16	14	14	13	17	2	8.00
20	21	22	22	21	25	19	24	3	12.00
21	16	15	15	14	14	14	16	1	7.50
22	10	7	6	7	8	5	5	2	4.25
23	14	10	8	10	10	8	7	3	6.25
24	23	13	9	6	14	8	11	7	7.50
25	25	16	16	16	17	16	16	0	11.50
26	8	10	10	10	10	9	10	1	10.00
27	17	18	18	18	18	18	19	0	9.75
28	25	8	7	4	9	3	7	5	4.00
29	20	6	6	6	7	7	6	0	5.00
30	17	13	12	13	13	11	10	3	4.75
31	26	8	8	8	8	9	8	0	5.50
32	28	7	7	6	5	5	5	2	2.50
33	25	12	12	8	11	8	9	4	4.50
34	1	2	2	2	2	2	2	0	1.50
35	31	6	6	6	6	6	6	0	5.75
36	25	11	10	11	10	11	10	1	4.50
37	26	17	15	8	8	8	15	9	5.50
38	28	14	14	14	15	15	13	1	7.75
39	15	11	11	11	11	11	11	0	4.50
40	18	13	13	7	11	8	9	6	4.25
41	10	16	16	15	16	18	16	1	7.50
42	28	21	18	21	16	18	17	5	7.75
43	15	10	10	10	10	10	10	0	6.50
44	30	7	7	6	4	4	6	3	2.25
45	29	12	10	11	10	9	17	3	2.75
46	33	4	2	4	2	2	13	2	0.75
47	19	13	10	10	16	10	10	3	3.75
48	16	16	17	16	16	16	17	0	5.50
49	17	16	15	16	16	16	17	1	5.75
50	32	9	9	10	10	8	10	1	4.50
51	25	10	10	9	10	9	9	1	5.50
52	13	12	12	13	13	13	13	0	9.00
53	23	13	13	14	12	14	12	1	6.25

Table 13: Incorrect Predictions, Combined Estimates, Repetition 2

Subject	Repetition							maximum	average
Number	RN	EU	DA	PR	RP	RQ	WU	difference	inconsistency
1	32	4	4	2	4	3	4	2	2.25
2	26	10	8	7	8	6	10	4	5.00
3	24	8	7	7	9	6	7	2	4.75
4	12	16	16	16	16	15	14	2	4.50
5	27	9	9	9	9	7	9	2	4.50
6	10	14	14	16	14	16	16	0	7.25
7	32	9	9	10	9	9	9	0	6.25
8	18	14	15	13	13	12	14	2	8.00
9	22	9	10	10	11	11	10	-1	6.00
10	19	12	12	12	10	12	10	2	7.00
11	27	5	5	5	4	4	5	1	3.00
12	30	1	1	1	1	1	1	0	2.00
13	8	19	19	18	12	17	14	7	6.50
14	12	12	15	13	12	14	11	1	6.75
15	11	11	12	10	11	9	9	2	7.00
16	22	8	9	5	8	5	6	3	5.75
17	14	11	11	11	13	12	14	0	9.50
18	18	10	12	12	10	13	12	0	5.75
19	20	16	15	15	15	14	16	2	8.00
20	24	17	19	20	18	18	21	-1	12.00
21	10	10	10	9	11	9	9	1	7.50
22	10	10	9	10	9	8	10	2	4.25
23	12	7	5	7	7	5	4	3	6.25
24	21	10	8	11	11	7	6	4	7.50
25	23	16	16	18	17	18	14	2	11.50
26	7	16	16	16	12	15	16	4	10.00
27	11	12	12	12	14	12	13	0	9.75
28	27	5	4	3	6	0	4	5	4.00
29	21	7	7	5	6	6	7	2	5.00
30	19	16	11	10	16	8	11	8	4.75
31	21	9	9	9	9	10	9	0	5.50
32	27	7	7	4	3	5	5	4	2.50
33	21	10	12	8	9	8	9	2	4.50
34	2	2	2	2	2	2	2	0	1.50
35	30	4	4	4	4	4	4	0	5.75
36	25	2	3	2	3	4	3	0	4.50
37	33	17	15	10	12	8	15	9	5.50
38	29	18	14	18	17	17	15	4	7.75
39	13	11	11	9	11	9	9	2	4.50
40	15	9	9	9	9	12	13	0	4.25
41	11	9	9	10	9	11	9	0	7.50
42	24	12	9	12	13	9	10	3	7.75
43	14	9	9	9	9	9	9	0	6.50
44	28	5	5	6	2	4	6	3	2.25
45	32	5	3	4	5	2	14	3	2.75
46	36	3	1	3	1	1	12	2	0.75
47	17	18	13	11	21	13	13	7	3.75
48	25	4	3	4	4	4	5	1	5.50
49	18	10	11	10	10	10	11	0	5.75
50	32	8	8	9	5	5	9	3	4.50
51	26	10	10	9	10	9	9	1	5.50
52	12	17	17	14	18	14	16	3	9.00
53	25	7	7	6	4	6	6	3	6.25

Table 14: Incorrect Predictions, Combined Estimates, Repetition 3

Subject	Repetition							maximum	average
Number	RN	EU	DA	PR	RP	RQ	WU	difference	inconsistency
1	32	3	3	1	3	2	3	2	2.25
2	26	9	9	8	9	7	9	2	5.00
3	26	7	6	8	6	5	6	2	4.75
4	10	10	10	10	10	11	8	2	4.50
5	27	9	7	9	7	7	9	2	4.50
6	8	13	13	9	11	11	11	4	7.25
7	33	6	8	7	8	6	6	0	6.25
8	16	20	21	17	17	18	16	4	8.00
9	25	11	10	10	11	11	10	1	6.00
10	21	8	8	8	8	8	6	2	7.00
11	29	6	6	6	3	5	6	3	3.00
12	29	3	3	3	3	3	3	0	2.00
13	6	10	10	7	9	8	9	3	6.50
14	18	13	12	12	13	11	12	2	6.75
15	15	13	12	14	13	15	9	4	7.00
16	25	9	6	10	9	10	7	3	5.75
17	14	13	13	13	7	12	8	6	9.50
18	20	8	10	10	8	11	8	0	5.75
19	18	14	11	9	13	10	10	5	8.00
20	31	13	9	16	12	12	13	4	12.00
21	10	10	10	11	11	11	11	0	7.50
22	12	12	13	12	9	12	12	3	4.25
23	11	11	7	9	11	7	10	4	6.25
24	25	12	10	15	11	13	14	2	7.50
25	17	19	17	19	16	17	17	3	11.50
26	8	13	13	13	13	14	15	0	10.00
27	11	16	16	16	18	16	13	3	9.75
28	29	12	11	8	13	7	11	5	4.00
29	26	9	9	7	6	6	9	3	5.00
30	21	11	10	11	11	11	10	1	4.75
31	33	5	5	5	5	6	5	0	5.50
32	28	5	5	4	3	5	5	2	2.50
33	24	5	7	5	6	5	6	0	4.50
34	1	1	1	1	1	1	1	0	1.50
35	28	7	7	7	9	7	7	0	5.75
36	27	4	3	4	3	4	3	1	4.50
37	32	8	6	7	7	5	6	3	5.50
38	22	12	8	12	11	11	7	5	7.75
39	15	7	7	5	7	5	5	2	4.50
40	17	9	9	7	5	8	9	4	4.25
41	10	13	13	12	13	13	11	2	7.50
42	13	13	14	9	20	14	15	4	7.75
43	16	7	7	7	7	7	7	0	6.50
44	28	6	6	5	3	1	5	5	2.25
45	33	4	2	3	4	1	15	3	2.75
46	35	2	0	2	0	0	11	2	0.75
47	17	19	14	14	22	14	10	9	3.75
48	30	7	8	7	7	7	6	1	5.50
49	24	10	9	10	10	10	9	1	5.75
50	28	4	6	5	7	7	5	-1	4.50
51	21	5	5	6	5	6	6	0	5.50
52	14	15	15	8	14	8	12	7	9.00
53	26	10	10	9	7	9	7	3	6.25

Table 15: Incorrect Predictions, Combined Estimates, Repetition 4

Subject	Repetition							maximum	average
Number	RN	EU	DA	PR	RP	RQ	WU	difference	inconsistency
1	32	3	3	3	3	2	3	1	2.25
2	28	10	10	9	10	8	10	2	5.00
3	29	8	7	9	5	6	7	3	4.75
4	14	11	11	11	11	10	13	1	4.50
5	30	7	5	7	5	5	7	2	4.50
6	11	16	16	10	14	10	12	6	7.25
7	33	3	5	4	5	3	3	0	6.25
8	13	15	16	14	14	13	11	4	8.00
9	23	9	8	8	9	9	8	1	6.00
10	23	6	6	6	6	6	4	2	7.00
11	29	7	7	7	4	6	7	3	3.00
12	29	0	0	0	0	0	0	0	2.00
13	4	15	15	12	12	11	10	5	6.50
14	15	14	15	13	14	14	13	1	6.75
15	17	9	10	12	9	13	7	2	7.00
16	17	14	13	11	14	11	12	3	5.75
17	16	17	17	15	15	14	14	3	9.50
18	20	8	6	12	6	13	8	2	5.75
19	11	9	10	12	10	11	7	2	8.00
20	30	15	13	16	12	12	11	4	12.00
21	11	12	12	13	11	13	13	1	7.50
22	6	9	10	9	8	9	11	1	4.25
23	11	12	12	10	12	8	11	4	6.25
24	20	24	20	17	25	19	22	7	7.50
25	21	20	18	20	19	20	18	2	11.50
26	18	8	8	8	8	9	8	0	10.00
27	13	17	17	17	19	17	14	3	9.75
28	29	5	4	3	6	4	4	2	4.00
29	30	9	9	9	10	10	9	0	5.00
30	26	10	7	8	10	8	11	3	4.75
31	30	8	8	8	8	9	8	0	5.50
32	27	4	4	3	2	4	4	2	2.50
33	27	7	7	9	8	9	6	1	4.50
34	1	2	2	2	2	2	2	0	1.50
35	25	10	10	10	12	10	10	0	5.75
36	29	3	2	3	2	3	2	1	4.50
37	33	11	9	6	8	6	9	5	5.50
38	24	10	10	10	11	11	11	0	7.75
39	14	6	6	4	6	4	4	2	4.50
40	21	11	11	11	9	12	11	2	4.25
41	12	16	16	15	16	16	12	4	7.50
42	21	11	8	15	12	10	5	6	7.75
43	18	12	12	12	12	12	12	0	6.50
44	29	6	6	5	3	3	5	3	2.25
45	34	4	2	3	4	3	15	2	2.75
46	35	2	0	2	0	0	11	2	0.75
47	15	18	15	15	21	15	11	7	3.75
48	28	5	6	5	5	5	4	1	5.50
49	22	6	7	6	6	6	5	1	5.75
50	27	13	11	12	16	10	14	3	4.50
51	18	9	9	10	9	10	10	0	5.50
52	14	15	15	10	14	10	12	5	9.00
53	26	14	14	13	11	13	11	3	6.25

Table 16: Incorrect Predictions, Combined Estimates, Repetition 5

Subject	Repetition							maximum	average
Number	RN	EU	DA	PR	RP	RQ	WU	difference	inconsistency
1	33	2	2	2	2	1	2	1	2.25
2	28	11	11	10	11	9	11	2	5.00
3	26	6	5	7	5	4	5	2	4.75
4	14	12	12	12	12	13	12	0	4.50
5	28	3	3	3	3	3	3	0	4.50
6	6	12	12	16	12	14	16	0	7.25
7	32	6	4	7	4	6	6	2	6.25
8	12	17	18	14	14	15	15	3	8.00
9	27	5	4	4	5	5	6	1	6.00
10	20	10	10	10	10	10	8	2	7.00
11	29	7	7	7	4	6	7	3	3.00
12	30	2	2	2	2	2	2	0	2.00
13	10	15	15	12	12	11	14	4	6.50
14	19	17	16	14	17	15	18	3	6.75
15	10	8	9	9	8	10	4	4	7.00
16	20	16	15	9	16	9	12	7	5.75
17	18	22	22	20	24	21	21	2	9.50
18	21	4	6	8	6	9	6	-2	5.75
19	24	14	13	13	13	12	16	2	8.00
20	25	15	15	18	16	16	17	0	12.00
21	11	6	6	7	7	7	7	0	7.50
22	10	10	11	10	9	10	8	2	4.25
23	13	8	8	8	8	8	7	1	6.25
24	19	15	13	14	18	16	15	2	7.50
25	26	15	13	13	12	13	13	3	11.50
26	25	17	17	17	19	16	15	2	10.00
27	8	19	19	19	19	19	14	5	9.75
28	25	6	5	8	7	5	5	1	4.00
29	28	8	8	8	7	7	8	1	5.00
30	23	11	8	9	11	9	12	3	4.75
31	28	5	5	5	5	4	5	1	5.50
32	28	6	6	5	6	6	6	1	2.50
33	24	9	9	7	10	7	8	2	4.50
34	1	1	1	1	1	1	1	0	1.50
35	25	9	9	9	9	9	9	0	5.75
36	30	6	5	6	5	4	5	2	4.50
37	32	13	11	8	10	8	13	5	5.50
38	26	7	5	7	6	6	6	2	7.75
39	16	6	6	4	6	4	4	2	4.50
40	19	9	9	5	7	6	7	4	4.25
41	12	8	8	7	8	8	4	4	7.50
42	20	11	12	15	14	14	9	2	7.75
43	24	6	6	6	6	6	6	0	6.50
44	30	8	8	7	5	3	7	5	2.25
45	36	3	1	2	3	2	16	2	2.75
46	35	2	0	2	0	0	11	2	0.75
47	14	19	16	12	22	16	10	9	3.75
48	30	7	8	7	7	7	6	1	5.50
49	22	2	3	2	2	2	1	1	5.75
50	31	14	12	11	13	5	13	9	4.50
51	21	7	7	8	7	8	8	0	5.50
52	9	13	13	12	14	12	12	1	9.00
53	19	9	9	10	10	10	10	0	6.25