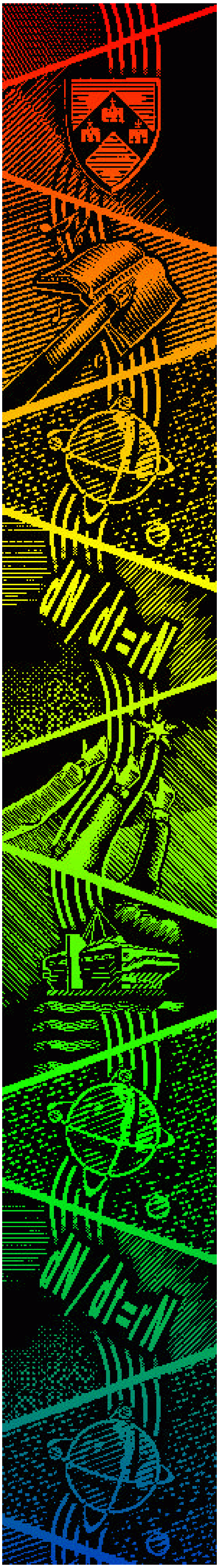

THE UNIVERSITY of York

Discussion Papers in Economics

No. 1999/18

Comparing Theories: What are we Looking For?

by

John Hey

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

Comparing Theories: What are We Looking For?

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

June 24, 1999

Abstract

Two recent papers, (Harless and Camerer, 1994) and (Hey and Orme, 1994) were both addressed to the same question: which is the 'best' theory of decision making under risk? The two papers shared a common concern: the appropriate trade-off between the descriptive accuracy of a theory and the predictive parsimony of that theory. In other respects, however, the two papers differed markedly: first in their treatment of the stochastic specification underlying the data generating process; second, and more importantly, in their interpretation of the question posed. This current paper tackles these two issues; first, trying to resolve the issue of the correct stochastic specification; second, by clarifying what economists might mean by a 'best' theory. The paper provides a general framework for answering such questions, and illustrates the application of this framework through two experiments aimed at answering the question: 'which is the best theory of decision making under risk?'.

1 Introduction

Two recent papers, (Harless and Camerer, 1994) and (Hey and Orme, 1994), were both addressed to the same question: which is the 'best' theory of decision making under risk? A second question that both addressed was: are any of the new generalisations of Expected Utility theory (EU) *significantly better* than EU (in some appropriate sense)? These are important questions: much theoretical effort has been expended in trying to produce a 'better' story of decision making under risk than that apparently provided by EU. What has been the purpose of this effort? Surely to improve the *predictive power* and *descriptive validity* of economics. These, of course, are competing objectives in general: other things being equal, the greater the predictive power of a theory, the lower the descriptive validity

[*] I am grateful to a number of people whose thoughts and ideas have influenced the development of this paper, particularly Bob Sugden and nrica Carbone.

of that theory. However, the purpose of 'better' theory is to make other things *not* equal. Nevertheless, there generally (and as it happens in the context of recent theories of decision making under risk, specifically) is the need to make some judgement of the appropriate trade-off between predictive power and descriptive validity: simply because if one theory was better in *both* predictive power and descriptive ability than a second, the second would simply be discarded - it would be dominated by the first.

Unfortunately, discarding dominated theories does not lead - in the area of decision making under risk - to a uniquely dominating theory. To discriminate amongst the remaining theories one therefore needs to do three things:

1. Decide on an appropriate measure of the predictive success of any theory

2. Decide on an appropriate measure of the predictive power of a theory

3. Decide on an appropriate way of trading-off the one against the other.

Selten (Selten, 1991) gives one possible set of answers to these questions; two recent papers (Harless and Camerer, 1994) and (Hey and Orme, 1994) give two interpretations of another. The purpose of this present paper is to try and shed light on their relative merits, as well as providing a general framework for the analysis of such questions.

Selten (Selten, 1991) suggests:

1. that we measure the predictive success of a theory as *the proportion of observations* in some given data set *consistent with that theory*

2. that we measure the predictive power of the theory by the *proportion of all possible observations* on that same data set *that are consistent with* (or predicted by) *that theory*

3. that the appropriate trade-off is simply given by the difference between these two proportions.

An illustration and application is given in (Hey, 1998). The main problem with this approach is that it leaves unresolved the key issue of the question of the meaning and interpretation of those observations *inconsistent* with the theory. Observations *consistent* with the theory are easy to interpret; but observations inconsistent with a theory are not so easy. A hardline approach requires us to interpret such observations as *refutations* of the theory - if we observe something inconsistent with a theory then that theory must be wrong. Unfortunately, if we proceed on this basis then we must conclude that *all* theories are wrong - since none predict *all* the observations on any given data set (unless we restrict the data set enormously). Selten's approach recognises this and therefore does not give a theory a rating of minus infinity if any inconsistent observations are noted; instead it treats all observations consistent with a theory the same (positive) weight and all observations inconsistent with a theory the same (finite and negative) weight. As Selten remarks:"A hit is a hit and a miss is a miss". I am not sure that all would agree. For instance, suppose on a set of 10 Pairwise Choice questions, that the only responses consistent with Expected Utility theory are *either* all Left *or* all Right, then according to Selten both 'LLLLLLLLLL' and 'RRRRRRRRRR' are consistent with EU, whilst anything else, for example 'LRRRRRRRRR' and 'LRLRLRLRLR' are inconsistent with EU. However, many others would want to qualify this, saying that 'LRRRRRRRRR' is somehow *nearer* to EU than is 'LRLRLRLRLR'. Selten's measure does not allow such discrimination. In contrast the approach used by (Harless and Camerer, 1994) and (Hey and Orme, 1994) does.

A further disagreement might be over Selten's suggested measure of the predictive power of a theory - which is effectively measuring what might be termed the *parsimony* of the theory. Is this really measured by the *proportion of the possible observations* on that same data set *that are consistent with* (or predicted by) *that theory*? As I shall argue, this depends upon what we are going to use our measure for - in other words, upon what we are going to use our analysis of the comparative ranking of the various theories for. Presumably this depends upon the application on which we are going to employ our 'best' or 'better' theories. It also depends upon the way that we are going to 'fit' our data to the various theories. As I shall show, the Selten measure of parsimony is very close to that used by Harless and Camerer - and this, in turn, is related to the way that they fit the data to the theories. Let me look at these two lines of argument in detail, beginning with the use to which we are going to put our analysis.

This all depends on the way we 'do' economics. If the economics we are 'doing' is a straight exercise in theory then one makes some assumptions about the objective functions of the various economic agents and then one explores the implications. Whether the theorist assumes the decision makers are EU maximisers or whether they are assumed to have some other objective function is in some sense irrelevant to what the theorist is doing - since the theorist can be argued to be simply exploring the implications of certain assumptions. So for the purpose of the exercise of straight economic theory the question of which is the 'best' theory of decision making under risk is irrelevant. But, of course, the exercise of straight theory is not the ultimate objective of economics - that must surely be the prediction of economic behaviour in a variety of contexts. Here we use the theory that the theorists have developed. But the way we use it must depend upon the context: we make assumptions about the economic agents in the context under study and then employ the relevant theory. We might then investigate whether the assumptions are valid and whether we might employ alternative or stronger assumptions. Clearly, in general, the stronger assumptions that we make the stronger the predictions that we can make - though, at the same time it is equally clear that the stronger the assumptions we make the more likely it is that these assumptions are incorrect. So we collect some relevant information about the particular context in which we are interested. For example, when predicting demand, we assume a particular form for the consumers' utility function(s), test whether that particular form appears to be consistent with the data, and (if relevant, which it almost always is) estimate any relevant parameters. *Occasionally* we may be able (or may have) to predict without any data at all, but such circumstances are unusual.

The context will determine what exactly it is that we are trying to predict - usually the aggregate behaviour of a group of individuals. However, given current economic methodology, much of microeconomic theory is a story about *individual* behaviour, so one needs to decide how one is going to solve the aggregation problem. Is it better to think of the group as represented by some representative individual and hence predict on the basis of that representative individual? Or is it better to work on the assumption that different people within the group are different, try to discover how many people are of each possible type and predict on the basis of such a characterisation? In general the second of these two approaches will work better if indeed different people in the group are different, though it could be the case that aggregation over the individuals averages out the individual responses in such a way that the aggregate looks as if it is behaving as though all the individuals were of a particular type. But the conditions for this to be so are likely to be strong - though much depends upon the context. Indeed there are contexts where the 'representative agent' model must be doomed to failure unless all people are identical: for example, consider the problem of predicting a group's choice in a pairwise choice problem (given information about the group's choices on earlier pairwise choice problems): the 'representative agent' model must necessarily predict that all the group would choose one or the other choice, whereas if there is a distribution of types, some will choose one option, others will choose the other[1].

These two different interpretations lead to two different ways of assessing how well various models fit the data. Of course, if the data set consists solely of aggregate data then there is no alternative but to fit the models to the aggregate data. But if one has individual data then one can implement both approaches. Let us suppose that that is the case. One wants to see how well the various theories fit the data. Occasionally a theory has no parameters - Expected Value Maximisation is an example of this - in which case there is no fitting to be done (unless one needs to estimate some error parameter). With other theories parameters are involved - which means that the appropriate parameters need to be chosen in some fashion to fit the theory to the data. Consider, as an example, the case of Expected Utility theory - which posits the maximisation of the expected value of

[1] Unless, of course, there is some stochastic element in behaviour. On this, see later.

some utility function. Unless one assumes that all agents are identical - and thus have the same utility function - then the 'parameters' that need to be chosen are the parameters that define the utility functions over the relevant domain. This can be done in general - or it could be done in a number of ways specifically for the data set under consideration. In order to explain what I mean by this, I need to give a specific example. This can obviously be generalised but it is difficult to make my point in a general context.

Suppose for example that the data set at hand is the set of responses of a set of I individuals, $i = 1, ..., I$, to a series of J pairwise choice questions, $j = 1, .., J$. Let the choice on question j by individual i be denoted by C_{ij} and suppose this can take one of two values L_j or R_j. The data set, therefore, consists of the ixj matrix $\mathbf{C} = C_{ij}, i = 1, .., I, j = 1, ..J$. Suppose further that individual i is an Expected Utility maximiser with utility function $u_i(.)$, then the individual's responses to the J questions can be either described by the value of $u_i(.)$ at the set of outcomes involved in the J pairwise choice questions, or the actual set of responses by the individual on the J questions. Note that the former imply the latter but the converse is not true. One could therefore argue that the former characterisation is more primitive in some appropriate sense.

Suppose, in addition to the data, one has a set of theories each of which is an attempt to explain the data. How might one fit the data to this set of theories? In general there are lots of ways of doing this - depending upon what restrictions, or assumptions, one imposes on the fitting process. Clearly the fewer the restrictions one places on the fitting process, the better that the fit is likely to be but the more 'parameters' one needs to estimate. Thus, if one is going to penalise the 'goodness of fit' of the data to the set of theories for the number of 'parameters' involved in the fitting, those fits with fewer restrictions are going to be penalised more heavily. One has a classic trade-off problem - which cannot be resolved in general but only in specific cases.

2 Ways of Fitting the Data to the Set of Theories

Let me list a partial set of the ways that one may 'fit' the data to the set of theories:

S1. One can assume that the behaviour of all agents in the data set is consistent with one particular theory (for example, Expected Utility theory) and that they all have exactly the same preference function (for example, in the case of Expected Utility theory, they all have the same (Neumann-Morgenstern) utility function.

S2. One can assume that the behaviour of all agents in the data set is consistent with one particular theory (for example, Expected Utility theory) but that different agents (potentially) have different preference functions (for example, in the case of EU theory, different agents (potentially) have different (Neumann-Morgenstern) utility functions).

S3. One can assume that different agents behave in accordance with different theories but that all those whose behavior is consistent with one particular theory share the same preference function relevant for that theory.

S4. One can assume that different agents behave in accordance with different theories and that agents whose behaviour is consistent with one particular theory may have differing preference functions (relevant for that theory).

As an empirical fact, one quickly discovers that, however few restrictions one imposes on the fitting method (unless the restrictions are so few that the whole exercise becomes meaningless), one is *unable to fit the data exactly.* What does one do? The obvious response - both for the economist and the econometrician - is to incorporate some story of *errors* into the fitting process. In the context of the majority of the currently popular theories of decision making under risk, this 'error' or noise term can very readily be interpreted as genuine error on the part of the decision maker[2]. So one needs a story of these errors - or at least, a stochastic specification of the errors. As I shall demonstrate, the choice of error story may limit what one can do in terms of fitting the data to the set of theories.

[2] There are theories of stochastic preference, see (Loomes and Sugden, 1995) and (Carbone, 1997a) and of stochastic choice with deterministic preference, see (Hey and Carbone, 1995) but here I shall concentrate on the mainstream literature which is a story of deterministic choice and deterministic preference. In this story 'noise' must be error.

3 Error Specifications

Let me concentrate on the two error stories proposed in the papers cited above: (Harless and Camerer, 1994) and (Hey and Orme, 1994). The first of these papers simply assumes that there is a probability θ that the agent will make a mistake[3] on any pairwise choice question - and that *this probability does not depend upon the nature of the pairwise choice question* itself. One can go further, as Harless and Camerer (Harless and Camerer, 1994) do and assume that θ is constant across all questions and indeed across all subjects, but this, of course, is not necessary. Again this depends upon how many restrictions one wishes to impose on the fitting and upon the resulting effect upon the goodness of fit. But one could adopt any of the following:

CP1. There is a probability θ_{ij} that subject i makes a mistake on question j.

CP2. There is a probability θ_i that subject i makes a mistake on *each* question.

CP3. There is a probability θ_j that *each* subject makes a mistake on question j.

CP4. There is a probability θ that *each* subject makes a mistake on *each* question.

I ignore, for the time being, the issue of the *identifiability* of these various models, issues which could be very severe, particularly for the first of these. Let me call these error specifications, respectively, CP1, CP2, CP3 and CP4, where CP stands for Constant Probability.

The story proposed in (Hey and Orme, 1994) is quite different. It goes back to the primitive of the preference functional $V(.)$ implied by the theory: according to a theory with preferences given by $V(.)$, L_j is preferred to R_j if and only if $V(L_j) > V(R_j)$, that is, if and only if $V(L_j) - V(R_j) > 0$. However, to accommodate the empirical 'fact' that agents make errors when c or right bias in the agent's answers) and possibly reasonably acceptable to assume that it has a normal distribution (appealing to the Central Limit Theorem). The magnitude of the error variance σ^2 can therefore be taken as a measure of the magnitude of the error spread: the larger is σ the greater in general will be the measurement error. Originally, (Hey and Orme, 1994) assumed that σ^2 was not dependent on the specific pairwise choice question, and I shall continue to work with that as a maintained hypothesis[4]. Nevertheless, there are still a variety of formulations that one could adopt:

WN1. That for subject i on question j the error variance is σ_{ij}^2.

WN2. That for subject i on *each* question the error variance is σ_i^2.

WN3. That for *each* subject on question j the error variance is σ_j^2.

WN4. That for *each* subject on *each* question the error variance is σ.

Again I ignore, for the time being, the issue of identifiability. I call these error specifications WN1, WN2, WN3 and WN4, where WN stands for White Noise (papers which have explored this type of specification extensively include (Carbone and Hey, 1994) and (Carbone and Hey, 1995) in addition to earlier references).

4 Describing True Preferences

In *principle* one can fit any of the model specifications combined with any of the error specifications, though we see that sometimes this is not possible. Sometimes this is because of a type of identification problem. Partly this depends on how we intend to describe the 'true' preferences, as defined by the specific preference functionals specified by the theory or theories in question. Let me return to that specification and illustrate with the case of Expected Utility theory. A *particular* EU preference function is defined by the underlying Neumann-Morgenstern utility function. This *might* be describable by a particular functional form, for example, linear, or constant absolute risk averse, or constant relative risk averse or it might not. Of course, one can always fit using a particular restricted functional form

[4] Though see (Hey, 1995) which suggests that specifying it as dependent on the questions might well improve the fit.

and the resulting saving in numbers of parameters to estimate may compensate for the worsening in the goodness of fit. An alternative is to specify the function at all possible values of its argument - but there may well be an infinite number of these, most being unidentifiable in any particular context. The best one can hope for, given that the pairwise choice questions must have been defined over a particular set of final outcomes, is to fit the function at those outcomes. Suppose there are L of these final outcomes, $O_l, l = 1, ..., L$. Then, at best, one can fit the function by estimating the value of $U(O_l)$ at the L values. Let me call this specification of the underlying true preferences as the specification of the underlying *True Values*. Now, as I have remarked before, any set of L values for $U(O_l)$ implies a particular set of responses on the J questions - for example: $L_1 L_2 R_3 .... L_J$; let me call this specification of the underlying true preferences as the specification of the underlying *True Responses*. Of course, these will be context specific, but then so will be the set of underlying True Values. Note crucially that it does not follow that a different set of $U(O_l)$ implies a different set of responses on the J questions; that is, it does not follow that a different set of underlying True Values implies a different set of underlying True Responses: there may be several sets of $U(O_l)$ consistent with any given set of responses to the J questions. Of course, in the context of a particular set of questions, knowledge of the sets of $U(O_l)$ consistent with a given set of answers does not increase the amount of knowledge gained from that data set; it just seems that it does[5]. In other words knowledge of the underlying *True Values* does not imply any extra knowledge - in a particular context - to knowing the underlying *True Responses*.

The above discussion has assumed that agents do not make mistakes. The evidence, however, would appear to contradict this. Of course, if agents do make mistakes then the way we specify their true preferences, combined with the way that we specify that they make mistakes, now has crucial and important significance. Consider a particular pairwise

[5] An interesting question is whether one can use the information gained from a particular set of questions to predict choice in some choice problem outside the original data set. The answer is that one could if it were the case that all sets of underlying true values consistent with a given set of responses implied a particular response on the new choice problem. This is unlikely to be the case but if it were then the information about the new choice problem would also have been implicit in the original responses.

choice question and suppose that an agent's true preference is for L_j. Let $\mathbf{a}(L_j)$ denote the set of parameter values of the underlying true preference functional which would give this particular preference. The CP errror specifications would give that the probability of the agent choosing R_j as $\theta_{(ij)}$ *irrespective of the actual value of the parameters* within the set $\mathbf{a}(L_j)$. In contrast the WN error specifications would imply that the probability of the agent choosing R_j is dependent upon the particular value of the parameters (within, of course, the set $\mathbf{a}(L_j)$). This implies that one can not use the WN error specification combined with underlying true preferences specified through the underlying True *Responses* - the reason simply being that, under the WN approach, the probability of making a mistake depends upon the underlying True Values and not just upon the underlying True Responses. However, and in contrast, one *can* use the CP error specifications with the underlying true preferences specified through the underlying *True Values* - though the implication is, as I will demonstrate, that the data does not allow us to discriminate between all underlying true values consistent with the estimated underlying True Responses. The reason for this is that the CP error specification identifies first the underlying True Responses and hence secondly but not uniquely the underlying True Values (the lack of uniqueness stemming from the fact that there is a *set* of underlying True Values consistent with any given underlying True Responses).

Given that one cannot use the WN error specification with the underlying true preferences specified through the underlying True Responses, but that one *can* use the CP error specification with the underlying true preferences specified through the underlying True Values, one might well be tempted to ask the question: why specify underlying true preferences through the underlying True Responses? Is there any advantage to doing so? The answer is: not really, at least when one understands what is the implication. There are some savings in computational effort -but these simply reflect the nature of the problem. For example, when using the CP error specification with the underlying true preferences specified through the underlying True Values, one discovers that the likelihood function (the thing we are trying to maximise - see later) is a *step function* when graphed as a function of the underlying True Values. This simply reflects the fact that this error specification does not distinguish between all values of the underlying True Values - indeed it *cannot* distinguish between those which imply the same set of observed responses, but only between those which imply different observed responses. The fact that the likelihood function is a step function creates computational and econometric problems - but these simply reflect the essentially *economic* nature of the problem in the first instance. Hence the difference between specifying the underlying true preferences through the underlying True Values or through the underlying True Responses is essentially cosmetic. This eliminates one apparent difference between the two papers under examination (Harless and Camerer, 1994) and (Hey and Orme, 1994). I shall work with whichever is most convenient. However, the aggregation problem should be kept in mind: although several agents may have the same underlying True Responses they may well *not* have the same underlying True Values.

Notwithstanding these theoretical considerations it remains the case that these computational difficulties are sufficiently important to shape the nature of the test that I wish to undertake. Ideally, I want a data set on which I can implement several of the above specifications. The problem is in implementing the CP error specification on data sets in which the number of questions J is at all large. If one characterises the problem in terms of the underlying True Responses, there is an interesting problem in determining the composition of the set of responses consistent with any particular theory. I have discussed this elsewhere (Hey, 1998) and will not rehearse the arguments here. Suffice it to say that for J at all large the number of possible responses 2^J is extremely large and the identification of the subset consistent with any given theory becomes a difficult task - particularly if the number of underlying True Values is itself large. Of course, one can carry out the fitting in the latter space, but if one is using the CP error specification this requires finding the maximum of a step function in a high-dimensioned space. And there is no guarantee that the function (the likelihood function) is everywhere concave in some appropriate sense[6].

[6] There is also the problem that one does not know where the next step is going to be, nor the width of it, which means that one could well miss the maximum. Indeed, with the algorithms currently in use - I have elsewhere used a Simulated Annealing program written in GAUSS by .G. Tsionas - there is no guarantee that the maximum will be found.

There are also complications if one wants to fit across *all* subjects.

If one is to employ one of the specifications in which agents are assumed to be different (at least partially) then one needs a reasonable amount of data for each subject. That is, one requires J to be reasonably large. This conflicts with the requirement of the paragraph above. I compromised by carrying out an experiment with $J = 15$. I also carried out a complete ranking experiment. The next section gives the details. The idea was to fit using both the CP error specification and the WN error specification so that the two could be compared.

5 The Experiments

I undertook two experiments - a Pairwise Choice experiment with J reasonably large (to be precise $J = 15$) and a Complete Ranking experiment. The Complete Ranking experiment was linked to the Pairwise Choice experiment in a sense that will be described shortly - but they were otherwise carried out completely independently of each other. Both involved gambles involving three final outcomes, which for the moment I shall refer to as x_1, x_2 and x_3 where these are indexed in such a way[7] that $x_1 \prec x_2 \prec x_3$ where $\prec$ denotes 'less preferred than'. A specific risky prospect is now described by the three numbers p_1, p_2 and p_3 where p_i denotes the probability that the outcome will be x_i $(i = 1, 2, 3)$. Note, however, that these three numbers must sum to unity - which means that any risky prospect can be described by just two of these three numbers. Take p_1 and p_3 - respectively the probability of the worst outcome and the probability of the best outcome. Now employ the expositional device known as the Marschak-Machina Triangle - with p_3 on the vertical axis and p_1 on the horizontal axis. See Figure 1. Each point within the Triangle represents some risky prospect; each of those on one of the sides of the Triangle is a prospect involving just two of the three outcomes; and those at the vertices of the Triangle are certainties (involving just one of the three outcomes). The 11 prospects I used in the Complete Ranking experiment

[7] We actually used amounts of money increasing in magnitude, so we are assuming that all our subjects preferred more money to less.

are the 11 points labelled a through k on this triangle. It will be noted that they all involve probabilities which are multiples of one-quarter. This was for a number of reasons, not least that, given the way we displayed the risky choices (see the Appendix containing the instructions for the Complete Ranking experiment), the probabilities were immediately and obviously discernible. In the Pairwise Choice experiment I used the same 11 basic prospects and presented to the subjects all posssible pairs involving these 11 prospects subject to the proviso that neither prospect in the pair dominated (in the first-degree sense) the other. There were 15 such pairs: specifically $ac, hc, hg, hi, fi, dc, dg, di, df, dj, de, ki, kj, ke$ and be. The reason why I omitted pairs in which one prospect dominated the other was that previous experimental evidence suggested that subjects virtually never chose the dominated prospect - in which case such questions would be uninformative. As it happened I observed surprisingly frequent violations of dominance on the Complete Ranking experiment. This suggests that subjects avoid violating dominance when dominance is obvious, but not necessarily otherwise - a view that has been gaining credence recently.

The Pairwise Choice experiment, with the 15 pairwise choices noted above, was carried out at **EXEC**[8] in York in 1995. The three outcomes were $x_1 = £0$, $x_2 = £300$ and $x_3 = £500$. I tried to recruit 250 subjects (the publicity material mentioned this number) but in the end I managed to recruit just 222[9]. To motivate the subjects, I used the following payment mechanism: after all 222 subjects had completed the experiment, all 222 were invited to a lecture room at a particular time. Each subject had a numbered cloakroom ticket identifying them; these tickets were put in a box and one selected at random. The subject with that number came to the front of the lecture theatre and drew at random one number from the set of integers 1 through 15. That particular subject's earlier-stated preferrred choice on that particularly-numbered pairwise choice question was then played out for real - and the subject paid accordingly. As it happened the subject was paid £300[10]

[8] The Centre for xperimental conomics at the University of York

[9] In a sense this number is irrelevant (as long as one gets 'enough' subjects - whatever that means) as long as it does not affect the choice made by the subjects.

[10] For those interested in such things,the winning subject was one who had approached me at the beginning of the meeting - having found some other subject's cloakroom ticket and having the honesty to say so. Clearly there is a reward for honesty!

- if the outcome had been £0 then the whole procedure would have been repeated from the beginning[11].

The Complete Ranking experiment was carried out (with the permission and very helpful cooperation of the conference organisers to whom I am most grateful) at the Seventh World Congress of the Econometric Society in Tokyo, Japan, in 1995. In the participants' conference packs there was included a single sheet inviting them to participate in this experiment; this invitation is reproduced in the Appendix. Anyone wishing to participate in the experiment - which involved simply ranking in order of preference the 11 basic prospects - had to hand in their ranking at the beginning of a lecture session at which I gave one invited paper (and Vince Crawford another). The experiment was played out at the end of the two lectures. Specifically, one of the answers was picked at random; the person concerned came to the front of the lecture room; then two of the 11 prospects were drawn at random by this person - and the one highest in that person's previously-stated ranking was played out for real.

In this experiment, the outcomes were denominated in American dollars: $x_1 = \$0, x_2 = \$200$ and $x_3 = \$1000$. Again the technique was deliberately to use large amounts of money and to pay off just one subject; my previous caveats apply[12]. It should also be noted that the middle outcome in the Complete Ranking experiment was chosen much closer to the worst outcome than in the Pairwise Choice experiment; this was because we had seriously misjudged the degree of risk aversion displayed by the subjects in the York experiment.

[11] An extended footnote is necessary at this stage. First, we should admit that playing the whole procedure repeatedly until someone had won something, slightly distorts the incentive mechanism - but since a different subject would (almost certainly) be chosen on each repetition the distortion is very slight. Second, although we could argue that this payment mechanism does give a strong incentive for honest reporting, in that *if* a particular subject is chosen and *if* a particular question is selected, then that subject will want (*ex post*) to have given his or her true preference on that question, the incentives might not be so strong as viewed from an *ex ante* perspective - given that the chance of being selected is so low. But ultimately, of course, this is an empirical issue: it would be interesting to explore the relative efficiency of using this procedure, as compared with using payoffs of one-tenth of these but paying off 10 subjects, or using payoffs one-hundredth of these, but paying off 100 subjects.

[12] Again, for those who like to know such things: the winner was a Russian academic and his winnings were $1000 - equivalent to approximately twice his annual salary! Proof that there is a God?!

6 Analysing the Results

If I was to fit all four models (S1 through to S4) specified above in conjunction with all the eight error specifications discussed above (CP1 through to CP4 and WN1 through to WN4) I would have to fit 32 different models to the data. Many of these can be discarded however. See Table 1; the following numbers refer to the entries in that table.

1. First, given the data set consisting of the results of the two experiments described above, the rows CP1 and WN1, involving the fitting of a different error parameter (either θ or σ) for each subject and for each question, cannot be implemented - the parameters are not identifiable, since questions were not repeated.

2. Harless and Camerer ((Harless and Camerer, 1994) would argue that we should also exclude rows CP2 and WN2 since "...allowing error rates to be choice-dependent can lead to nonsensical results." (page 1261).

3. I would argue that we should exclude column S1 since the notion that all subjects in our experiment had exactly identical tastes is manifestly absurd.

4. I would also go further and exclude column S3 on the argument that if we are prepared to accept that different agents may have different preference functionals it is then odd to argue that all those with the same functional should also have the same tastes within that functional.

5. I would eliminate the remainder of the WN4 row on the grounds that the empirical evidence obtained from the estimation of the WN2 row is that the error variances clearly vary considerably from subject to subject.

6. Finally I would eliminate column S4 combined with row CP4: if subjects really are as different as implied by S4 it is highly unlikely that they are identical in the way indicated by CP4.

As far as columns are concerned this leaves us with two - S2 and S4, effectively the representative agent model and the varied agent model. A comparison of the fitting for the two columns enables us to see which of these two stories appears to be the better. Generally we are left with specifications A through E, as follows:

Specification A: [S2,CP2] All subjects have the same preference functional but different (CP) error parameters. This is particularly simple to fit: for each subject we find the 'nearest' set of consistent responses (consistent with a particular theory) to the observed responses (nearest in the sense of the smallest number of mistakes between the consistent responses and the observed responses). We then add up the log-likelihoods across all subjects, theory by theory, correct them for degrees of freedom (as described below) and choose that preference functional for which the corrected log-likelihood is maximised.

Specification B: [S4, CP2] Different subjects have different preference functionals and different (CP) error parameters. We follow the procedure described above, but then work subject by subject, rather than preference functional by preference functional: for each subject we find the preference functional for which the corrected log-likelihood is maximised (corrected in the manner described below) and then aggregate the corrected log-likelihoods over all subjects. Because the correction procedure is different from that in Specification A (see below) there is no guarantee that this Specification does worse or better than Specification A.

Specification C: [S2, CP4] This is the original Harless and Camerer specification: all subjects have the same preference functional and the (CP) error is constant across subjects. We calculate the log-likelihood *across all subjects*, preference functional by preference functional, correct them for degrees of freedom and then aggregate.

Specification D: [S2,WN2] All subjects have the same preference functional but they have different (WN) error parameters. This is similar to Specification A except that we use the WN error specification. We work preference functional by preference functional, aggregating the maximised log-likelihoods across all subjects, correcting them for degrees of freedom and then choose that preference functional for which the corrected log-likelihood is maximised. Because the correction factor is the same as in Specification E, this is *bound* to do no better than Specification E. Nevertheless, it is interesting to see how much worse it performs.

Specification : [S4, WN2] This is the original Hey and Orme specification: different subjects (may) have different preference functionals with differing (WN) error parameters. We follow the procedure described above, but then work subject by subject, rather than preference functional by preference functional: for each subject we find the preference functional for which the corrected log-likelihood is maximised (corrected in the manner described below) and then aggregate the corrected log-likelihoods over all subjects.

There are some interesting estimation problems involved with the CP stories: as described in the original paper (Harless and Camerer, 1994) the fitting problem is one of finding the proportion of subjects in the sample with underlying true responses of each type consistent with any one theory. This is the case when the error parameter θ is assumed to be constant across both questions and subjects. In this case, the interpretation as to what is implied for any particular subject is that one is estimating the probabilities that the subject's underlying true responses are each of the allowable ones: the overall fitted proportions are the weighted average of these probabilities, averaged over all observed responses. In contrast, when one assumes that the error parameter, θ_i, varies across subjects (but not across questions) then the maximum likelihood estimator of θ_i is the minimised proportion of mistakes (across all questions for that particular subject). So fitting this story is equivalent to finding, for each subject, the response consistent with the appropriate theory closest to the subject's actual response - closest in the sense of the smallest number of errors implied by the actual response if that consistent response were indeed the subject's underlying true response. In this case the maximised log- likelihood is simply the maximum of $ln[\theta_i^j(1-\theta_i)^{(J-j)}]$ where J is the total number of questions and j the number of incorrect responses given the underlying true consistent response. This maximised likelihood is achieved when $\theta_i = j/J$ and takes the value $jln(j) + (J-j)ln(J-j) - Jln(J)$.

7 Correcting for Degrees of Freedom

It is clear that different specifications involve different numbers of estimated parameters. Clearly also it is the case that the more parameters involved in the fitting of a particular specification, the better that specification will fit. Goodness of fit is measured by the maximised log-likelihood for that specification. One therefore needs a way of 'correcting' the maximised log-likelhood for the number of parameters involved in the fitting.

This is a familiar problem in econometrics; there are a number of recommended solutions - none obviously superior to all others. I therefore simply adopt one of the more familiar ones - namely the Aikake Criterion[13]. This involves maximising $2ln[L(\hat{\alpha})] - 2k/T$ where $L(\hat{\alpha})$ is the maximised likelihood, T the number of observations and k the number of parameters involved in the fitting. Given that, in the comparisons I will be carrying out, the number of observations T will be constant, this is equivalent to maximising $ln[L(\hat{\alpha})] \ - \ k$. In other words, we simply correct the maximised log-likelihood by subtracting from it the number of parameters involved in its fitting. Let me know turn to consideration of the number of parameters involved in the fitting of the various specifications.

I need two bits of notation. Denote by M_k the number of consistent responses under theory k. This obviously varies from theory to theory (preference functional to preference functional) and clearly also depends upon the specific questions asked in the experiment.

Let me also denote by N_k the number of underlying true values required under theory k. Again this will vary across theories and will depend upon the specific questions in the experiment. In the context of my two experiments - with just 3 outcomes - then N is *zero* for the Risk Neutral preference functional (as there are no parameters involved with it), N is *one* for the Expected Utility functional - since the utility of two of the three outcomes are normalised (to zero and unity) leaving just one utility value to be determined. As we shall see later, N is *two* for all the other theories under consideration - as the fitting involves just one utility value (as in Expected Utility therory) and one other parameter.

I can now specify the number of parameters involved with each specification and hence summarise my procedure for ranking and comparing the various sepcifications. Let LL^*_{ik} denote the maximised log-likelihood function for subject i on theory k if the specification allows us to fit subject by subject. If not, use LL^*_k to denote the maximised log-likelihood

[13] For a Monte-Carlo investigation of the efficiciency of this criterion, see (Carbone and Hey, 1994) and (Carbone, 1997b).

across all subjects. Then it works as follows:

Specification A: [S2,CP2] All subjects have the same preference functional but different (CP) error parameters. Then, for each preference functional, we estimate the proportion of subjects with each of the M_k true responses - thus giving us $(M_k - 1)$ parameters to estimate (because these M_k proportions must sum to one - and, *for each subject* i, estimate that subject's error parameter θ_i. We then choose that preference functional for which the following expression is maximised:

$\max_{k=1}^{K} \ (\sum_{i=1}^{I} \ [LL_{ik}^{*} - 1]) - [M_k - 1])$

Specification B: [S4, CP2] Different subjects have different preference functionals and different (CP) error parameters. Because we are now effectively fitting subject by subject it is better if we fit the N_k true values. We thus get as our maximised corrected log-likelihood:

$\sum_{i=1}^{I} \ \max_{k=1}^{K} \ [LL_{ik}^{*} - N_k - 1]$

Specification C: [S2, CP4] This is the original Harless and Camerer specification: all subjects have the same preference functional and the (CP) error is constant across subjects We are therefore fitting $M_k - 1$ proportions (the final one being determined by the fact that they must sum to unity) and one error parameter. We thus get:

$\max_{k=1}^{K} \ [LL_{k}^{*} - M_k]$

Specification D: [S2,WN2] All subjects have the same preference functional but they have different (WN) error parameters. So for each subject we need, for preference functional k, to fit N_k values and one error parameter σ_i. We thus get:

$\max_{k=1}^{K} \ \sum_{i=1}^{I} \ [LL_{ik}^{*} - N_k - 1]$

Specification : [S4, WN2] This is the original Hey and Orme specification: different subjects (may) have different preference functionals with differing (WN) error parameters. The story is the same as Specification D, though the aggregation and maximisation are done in reversed orders. Thus the expression below is bound to be higher than that for Specification D above. We have:

$\sum_{i=1}^{I} \ \max_{k=1}^{K} \ [LL_{ik}^{*} - N_k - 1]$

8 Full and Overfull Correction

There is one *caveat* that needs to be made to the above discussion: it assumes that different subjects respond differently. If, however, they do not not, then one could argue that the correction is excessive. If one has j subjects all with the same response, then under all specifications other than Specification C, one could argue that having fitted one of these j subjects then the other $j-1$ are also fitted by the same parameter values - one does need to repeat the correction. However, one does need to repeat the maximised log-likelihood as the other $j-1$ subjects are genuine observations. This is the procedure followed in the tables below: under the 'full correction' only one set of corrections is implemented for multiple (repeat) observations. The 'overfull corrections' carry out a correction for each subject, irrespective of whether they have the same experimental responses as other subjects. I would argue that the *Full Correction* is the correct procedure.

9 CP Errors in the Complete Ranking Experiment

Given that Harless and Camerer introduced their error story in the context of Pairwise Choice experiments, and given that, to the best of my knowledge, this story has not been extended to the Complete Ranking context, I must make the extension myself. Whilst I have consulted with David Harless over this, I cannot be sure that this meets with his approval.

Consider a ranking of two objects, and suppose the true ranking is '12'. If the subject states this, there is no error; if he or she instead reports '21', then there is one error. Consider now three objects, and suppose '123' is the true ranking. Then '132' or '213' could be considered as one mistake - just one item in the wrong position - and '321' could be considered two mistakes. Such considerations lead to the following story. Suppose there are Z objects to rank and suppose the true ranking is $x_1x_2...x_Z$ but the reported ranking is $y_1y_2...y_Z$ then one could argue that the 'number of mistakes' made is $\sum_{z=1}^{Z}|x_z - y_z|/2$. This is the measure I used. In keeping with the spirit of the CP approach I assumed that (under the CP specifications) the probability of making any one of such mistakes was a constant (independent of the context).

10 Preference Functionals Fitted

In addition to the models already discussed (Risk Neutrality and Expected Utility) I fitted five other 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**). Details of these can be found in (Hey, 1997). 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.

11 Results

Let me discuss the results specification by specification first. Begin with Specification A in Table 2. If one judges, as I have argued one should, on the basis of the Fully Corrected Log-Likelihood, then Prospective Reference theory (**pr**) emerges as the 'best' functional on the Pairwise Choice experiment, and Rank dependent with the Quiggin weighting function (**rq**) on the Complete Ranking experiment. This echoes earlier findings. Expected Utility theory does not do particularly well - as a Representative Agent model - and neither does Disappointment Aversion theory especially in the Complete Ranking experiment.

Specification B is summarised in Table 3. Details of the 'best' model are given in Table 4, which specifies the number of subjects for whom a particular model was 'best' in terms of the Corrected Log-Likelihood[14]. It may be of interest to note that Risk Neutrality comes best for 10 subjects on the PC experiment and best for 44 on the CR experiment. Corresponding figures for EU are 165 (PC) and 22 (CR), whilst a top-level functional (one

[14] When k models tied for 'best' under this criterion, each was given a score of $1/k$.

of **da**, **pr**, **rp** , **rq** or **wu**) came best for just 47 subjects on PC and 59 subjects on CR. (Recall there were 222 subjects on the PC experiment and 125 on the CR experiment.) Prospective Reference theory (**pr**) did particularly well on the PC experiment and the Rank Dependent models on the CR experiment. It is interesting to note that Specification B does marginally worse than Specification A on the Pairwise Choice experiment, though marginally better on the Complete Ranking experiment.

Specification C is summarised in Table 5. This is the original Harless and Camerer specification. It performs considerably worse than Specifications A and B - indicating that the constant-across-all-subjects error hypothesis looks highly suspect - as one might imagine. For the record, EU does 'best' for the PC experiment and Weighted Utility (**wu**) for the CR experiment. But one should not attach too much weight to these remarks.

Specification D is summarised in Table 6. Remember that this is bound to do worse than Specification E - but the difference is not too large. From Table 6 it can be seen that Prospective Reference theory (**pr**) does 'best' on the PC data and Weighted Utility on the CR data.

Finally, specification E is summarised in Table 7. The breakdown of 'best' models is summarised in Table 8. It can be seen that Risk Neutrality and Expected Utility theory do rather well.

An overall summary is provided in Table 9. It is particularly clear from this that Specification C (the original Harless and Camerer specification) does rather badly. The 'best' specification appears to be that of Specification E - the original Hey and Orme specification. I suspect that this is the combined incidence of two effect, first a possibly better error specification[15] and partly and perhaps more importantly, because Specification C embodies the Representative Agent model which seems to be seriously misleading[16]. The

[15] Though elsewhere (Carbone and Hey, 1997) I provide direct evidence to compare the WN error specification with the CP error specification, from which it is not clear that either can be regarded as *generally* superior.

[16] It may be interesting to 'translate' the maximised log-likelihoods into probabilities for individual subjects on individual questions. On the Pairwise Choice experiment the LL figure of -625 for Specification is equivalent to a probability on average of 0.829 on each question for each subject of observing what was observed given the fitted model. In contrast, the LL figure of -992 for Specification C is equivalent to a probability of 0.742.

evidence of this paper must surely be that people *are* different.

12 Conclusions

Two methods of assessing and comparing theories have been referred to in this paper: the Selten method and the Harless/Camerer/Hey/Orme (HCHO) method. Both penalise the ‘goodness of fit’[17] of theories through some measure of the parsimony of the theory. The Selten penalisation turns out to be effectively the same[18] as that of HCHO in the context of the Harless and Camerer method of fitting the data to the theories (Specification C). This penalisation is effectively the number of parameters involved with the fitting of the specification - and is familiar to econometricians. In other specifications it needs to be modified appropriately. But it is not *this* that distinguishes Selten from HCHO. Rather it is in the measurement of ‘goodness of fit’ or predictive success: Selten (“A miss is a miss and a hit is a hit”) counts all observations consistent with a theory as successes and all those inconsistent as failures. In contrast HCHO measure how bad misses are - near misses being better for a theory than distant misses. This requires a stochastic specification (which, of course, Selten’s does not) and allows the use of the Maximised Log-Likelihood as the measure of predictive success. The stochastic specification differs between Constant Probability and White Noise. A peripheral question answered in this paper concerns which of the two is empirically best, but the major finding is that one can view both Harless and Camerer and Hey and Orme as two attempts to answer the same question within the same basic framework. This paper has made clear what that framework is.

Fundamentally the issue at the heart of this paper boils down to the question of the best (corrected) fit - which is a essentially empirical question. As it happens, with the data set that we have, it appears to be the case that the Representative Agent model performs particularly badly - with the conclusion being that it is better to treat different people as different. Doing otherwise leads to worse predictions - notwithstanding the improved

[17] Here measured by the Maximised Log-Likelihood.

[18] Compare the penalisation used in this paper with that in (Hey, 1998).

parsimony.

And finally, as far as the ‘Best’ theory of decision making under risk is concerned, our analysis tells us that we should not discard Expected Utility theory. Nor should we discard *all* the many new theories - some are ‘best’ for some subjects - though there are some theories which look of increasingly minor interest.

References

Carbone, E. (1997a). Discriminating between preference functionals: A Monte Carlo study. *Journal of Risk and Uncertainty*, 15:29–54.

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

Carbone, E. and Hey, J. D. (1994). Discriminating between preference functionals - a preliminary Monte Carlo study. *Journal of Risk and Uncertainty*, 8:223–242.

Carbone, E. and Hey, J. D. (1995). A comparison of the estimates of EU and non-EU preference functionals using data from pairwise choice and complete ranking experiments. *Geneva Papers on Risk and Insurance Theory*, 21:111–133.

Carbone, E. and Hey, J. D. (1997). Which error theory is best? unpublished.

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

Hey, J. D. (1995). Experimental investigations of errors in decision-making under risk. *uropean conomic Review*, 39:633–640.

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

Hey, J. D. (1998). An application of Selten's measure of predictive success. *Mathematical Social Sciences*, 35:1–16.

Hey, J. D. and Carbone, E. (1995). Stochastic choice with deterministic preferences - an experimental investigation. *conomics Letters*, 47:161–167.

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

Loomes, G. C. and Sugden, R. (1995). Incorporating a stochastic element into decision theory. *uropean conomic Review*, 39:641–648.

Selten, R. (1991). Properties of a measure of predictive success. *Mathematical Social Sciences*, 21:153–167.

Table 1: Various possible specifications

Error/Model	Error parameter	S1	S2	S3	S4
CP1	θ_{ij}	13	1	14	1
CP2	θ_i	3	A	4	B
CP3	θ_j	23	2	24	2
CP4	θ	3	C	4	6
WN1	σ_{ij}	13	1	14	1
WN2	σ_i	3	D	4	E
WN3	σ_j	23	2	24	2
WN4	σ	3	5	4	5

Table 2: Log-Likelihoods for Specification A

Preference	Pairwise Choice			Complete Ranking		
Functional	Correction			Correction		
Fitted	None	Full	Overfull	None	Full	Overfull
rn	-2010	-2090	-2232	-1272	-1336	-1397
eu	-675	-760	-902	-848	-917	-978
da	-615	-723	-865	-690	-782	-843
pr	-578	-693	-835	-592	-679	-780
rp	-640	-766	-908	-556	-666	-727
rq	-584	-729	-871	-462	-591	-652
wu	-594	-721	-863	-519	-630	-691

Table 3: Log-Likelihoods for Specification B

Pairwise Choice			Complete Ranking		
Correction			Correction		
None	Full	Overfull	None	Full	Overfull
-554	-744	-1035	-353	-527	-618

Table 4: 'Best' Models under Specification B

Preference Functional	Pairwise Choice	Complete Ranking
rn	10.00	44.00
eu	165.00	22.00
da	4.50	4.67
pr	18.75	8.67
rp	4.92	17.00
rq	11.92	17.50
wu	6.92	11.17

Table 5: Log-Likelihoods for Specification C

Preference	Pairwise Choice			Complete Ranking		
Functional	Correction			Correction		
Fitted	None	Full	Overfull	None	Full	Overfull
rn	-2065	-2066	-2066	-1746	-1747	-1747
eu	-985	-992	-992	-1348	-1354	-1354
da	-982	-1010	-1010	-1237	-1266	-1266
pr	-977	-1011	-1011	-1167	-1231	-1231
rp	-972	-1019	-1019	-1140	-1187	-1187
rq	-973	-1039	-1039	-1057	-1123	-1123
wu	-976	-1024	-1024	-1115	-1161	-1163

Table 6: Log-Likelihoods for Specification D

Preference	Pairwise Choice			Complete Ranking		
Functional	Correction			Correction		
Fitted	None	Full	Overfull	None	Full	Overfull
rn	-2145	-2225	-2367	-963	-1027	-1088
eu	-613	-773	-1057	-408	-536	-658
da	-527	-767	-1193	-340	-532	-715
pr	-467	-707	-1133	-266	-458	-641
rp	-516	-756	-1182	-298	-490	-673
rq	-518	-758	-1184	-257	-449	-632
wu	-500	-740	-1166	-250	-442	-625

Table 7: Log-Likelihoods for Specification E

Pairwise Choice			Complete Ranking		
Correction			Correction		
None	Full	Overfull	None	Full	Overfull
-429	-625	-938	-200	-377	-466

Table 8: 'Best' Models under Specification E

Preference Functional	Pairwise Choice	Complete Ranking
rn	13	42
eu	131	25
da	13	9
pr	32	7
rp	6	16
rq	17	10
wu	10	16

Table 9: Overall Summary of Log-Likelihoods

Specification	Pairwise Choice			Complete Ranking		
	Correction			Correction		
	None	Full	Overfull	None	Full	Overfull
A	-578	-693	-835	-462	-591	-652
B	-554	-744	-1035	-353	-527	-618
C	-972	-992	-992	-1057	-1123	-1123
D	-467	-707	-1133	-250	-442	-625
E	-429	-625	-938	-200	-377	-466

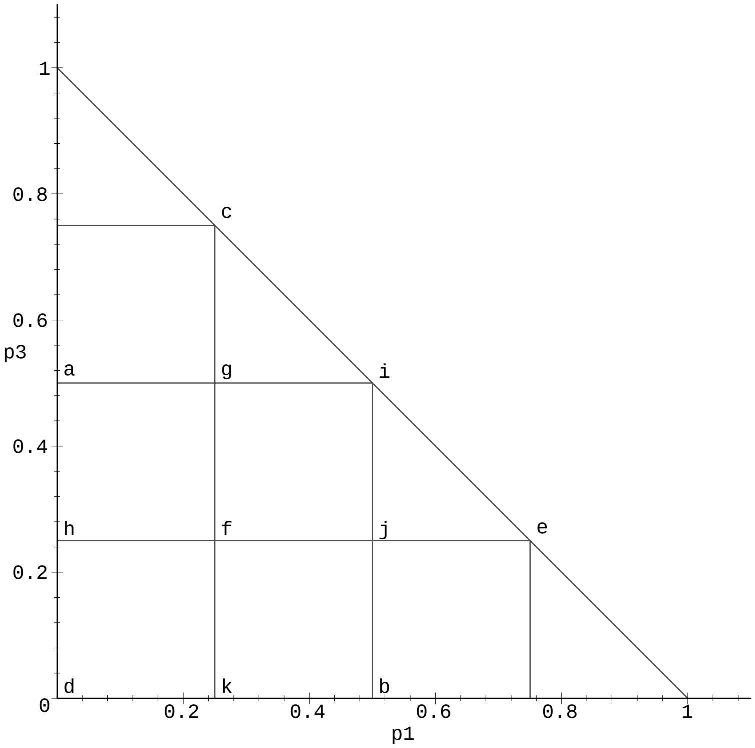

Figure 1: The Risky Choices in the two Experiments