On the theory of reference-dependent preferences*

Alistair Munro and Robert Sugden

School of Economic and Social Studies

University of East Anglia

Norwich NR4 7TJ, UK

March 2001

Correspondence should be sent to: Robert Sugden

JEL classification: D11 (consumer economics: theory), D51 (exchange and production economies)

Keywords: status quo bias, loss aversion, reference-dependence, prospect theory

* This paper grew out of research supported by the Economic and Social Research Council of the UK (award nos W 119 251 014 and L 211 252 053); Robert Sugden's work has also been supported by the Leverhulme Trust. We are grateful for comments from Ian Bateman, Indranil Dutta, Daniel Kahneman and Chris Starmer.

Abstract

A theory is proposed in which preferences are conditional on reference points. It is related to Tversky and Kahneman's reference-dependent preference theory, but is simpler and deviates less from conventional consumer theory. Preferences conditional on any given reference point satisfy standard assumptions. Apart from a continuity condition, the only additional restriction is to rule out cycles of pairwise choice. The theory is consistent with observations of status quo bias and related effects. Reference points are treated as subject to change during the course of trade. The implications of endogeneity of reference points for behaviour in markets are investigated.

There is now a great deal of evidence that, contrary to the assumptions of Hicksian consumer theory, individuals' preferences between given options vary systematically according to what is perceived to be the reference point – that is, the status quo, or the customary or normal state of affairs. It is particularly well-established that a person is more likely to prefer a given option to some alternative if that option is the status quo than if it is not: this is the *endowment effect* or *status quo bias*, first described by Richard Thaler (1980). But this is only one of a range of apparently related reference point effects. Surprisingly little work has been done to develop theories of consumer choice, appropriate for use in economics, which can accommodate these effects.

The most fully worked out such theory is probably Amos Tversky and Daniel Kahneman's (1991) theory of reference-dependent preferences, in which an individual's preferences over bundles of goods are conditional on his or her reference point. This theory successfully organises a range of regularities that have been found in experimental and survey research. However, it does so at the cost of some radical deviations from conventional consumer theory, not all of which improve the theory's predictive power. We shall suggest that these apparently redundant features of Tversky and Kahneman's theory are motivated by certain analogies between consumer choice and choice under risk. These depend on an implicit assumption that preferences are additively separable – a restrictive assumption which seems orthogonal to the analysis of reference point effects. In this paper, we propose a reformulation of Tversky and Kahneman's theory which, while equally consistent with the evidence on reference point effects, requires fewer departures from conventional consumer theory.

Tversky and Kahneman's theory treats reference points as exogenously given. This approach is adequate for explaining most of the experimental and survey data, but limits the economic applicability of the theory. The evidence suggests that individuals adjust their reference points remarkably quickly in response to changes in their endowments. Thus, acting in accordance with the preferences appropriate to one reference point, an individual may make trading or consumption decisions which then induce a change in the reference point; this change may lead to further trade or to revised consumption decisions, and so on. If we are to explain behaviour in markets in anything other than the very short run, we need a theory in which reference points are *endogenous*. The theory we propose has this property.

1. Reference point effects: the stylised facts

In reviewing the evidence of reference point effects, we focus on the domain of consumer theory: situations of certainty in which individuals choose between alternative multi-dimensional bundles of goods.

Before going on, we must acknowledge an ambiguity in the concepts of 'status quo' and 'reference point', as they are used in the literature of reference-dependent preferences. On one interpretation, a consumer's status quo position is given by his *current endowment* – the bundle of goods that he currently owns. On another interpretation, it is given by his *customary consumption* – the bundle of goods that he has become used to consuming. To see the difference between these interpretations, consider a consumer who is in the habit of buying a newspaper each morning. As he approaches the news-stand on a particular morning, is the status quo position that he buys the paper (his customary consumption) or that he does not buy it (his current endowment)? In their theoretical work, Tversky and Kahneman explicitly leave open the question of how reference points are determined; when they describe the evidence that supports their theory, they move freely between alternative interpretations (1991, pp. 1040-1045). We will discuss the significance of the distinction between endowments and customary consumption later in the paper. For the present, however, we use the term 'reference point' to encompass both interpretations.

The largest body of evidence of reference point effects concerns the *WTA/WTP discrepancy*. This discrepancy was first found in contingent valuation surveys which elicit willingness-to-pay (WTP) and willingness-to-accept (WTA) valuations of the same good. The early discovery – for example, by Richard Bishop and Thomas Heberlein (1979) and Robert Rowe, Ralph d'Arge and David Brookshire (1980) – that mean and median values of WTA are often several times higher than the corresponding WTP values has been replicated many times. The same discrepancy has been found in incentive-compatible laboratory experiments. Some of these experiments – for example, those of Jack Knetsch and John Sinden (1984) and of Ian Bateman et al (1997) – have used designs which control for the income and substitution effects which some commentators have used to explain WTA/WTP discrepancies in contingent valuation studies.

A closely related phenomenon is found when individuals choose between alternative options, one of which is perceived as the status quo. Behaviour shows a regularity which may be called

conservatism: for given options x and y, individuals are more likely to choose x in preference to y if x is the status quo than if y is. The classic experiment is that of Knetsch (1989). Subjects were divided at random into two groups; the members of one group were given coffee mugs, the members of the other were given chocolate bars. A few minutes later, each subject was given the option of exchanging the gift he had received for the other one; marked conservatism was observed. This effect has also been found outside the laboratory. For example, William Samuelson and Richard Zeckhauser (1988) find evidence of conservatism in the decisions of Harvard University employees in relation to medical and pension plans.

Experiments by Knetsch (1989) and by George Loewenstein and Daniel Adler (1995) have found that individuals who choose between an increment of money and some specific good are more likely to choose the money if the decision problem is presented to them as a choice between alternative gains than if they are endowed with the good and are then invited to exchange it for the money. A similar effect has been found for valuations. An individual reports an *equivalent gain* (EG) valuation by reporting the smallest amount of money that he would accept in place of a given increase in some specific good. Similarly, an individual reports an *equivalent loss* (EL) valuation by reporting the largest money loss that he would accept in place of a given decrease in the specific good. Kahneman, Knetsch and Thaler (1990) find that EG valuations lie between WTP and WTA: we shall call this the *EG effect*. Bateman et al (1997) replicate this finding, and find that EL valuations also lie between WTP and WTA (the *EL effect*).

To summarise the evidence, we use Figure 1. Consider any two bundles of two goods, neither of which dominates the other. Let these bundles be $x = (x_1, x_2)$ and $y = (y_1, y_2)$, where $y_1 > x_1$ and $x_2 > y_2$. For any given individual, we can investigate whether his (revealed or reported) preference ranking of x and y varies as the reference point changes from one point in commodity space $r^i = (r^i_1, r^i_2)$ to another point $r^j = (r^j_1, r^j_2)$. We shall say that such a change in the reference point *shifts preferences towards* y if, in between-individual comparisons, y is more likely to be preferred to x when the reference point is r^j than when it is r^i. All such shifts in preference imply contraventions of Hicksian consumer theory.

The WTA/WTP discrepancy and conservatism both imply that a change in the reference point from x to y shifts preferences towards y. The EG effect implies that this change can be broken down into two components: preferences are shifted towards y *both* if the reference point changes

from x to $r^5 = (x_1, y_2)$ *and* if it changes from r^5 to y. The EL effect implies that another such decomposition is possible: preferences are shifted towards y *both* if the reference point changes from x to $r^7 = (x_2, y_1)$ *and* if it changes from r^7 to y.

Some evidence exists about the effects of other changes in the reference point. First, compare the reference points r^4 and r^6, defined by $r^4_1 = x_1$, $x_2 > r^4_2 > y_2$, $y_1 > r^6_1 > x_1$, and $s^6_2 = y_2$. Notice that r^4 is dominated by x but not by y, while r^6 is dominated by y but not by x. Tversky and Kahneman (1991, pp. 1044-1045) and Kaisa Herne (1998) report that changes in the reference point from r^4 to r^6 shift preferences towards y. Next, compare r^3 and r^8, defined by $x_1 > r^3_1$, $r^3_2 = x_2$, $r^8_1 = y_1$, and $y_2 > r^8_2$. Notice that r^3 is dominated by x but not by y, while r^8 is dominated by y but not by x. Herne finds that changes in the reference point from r^3 to r^8 shift preferences towards y. She classifies this and the preceding effect as *asymmetric dominance effects.* To distinguish them, we shall call the first effect the *inner* and the second the *outer* asymmetric dominance effect.

Now compare r^1 and r^{10}, defined by $x_1 > r^1_1$, $r^1_2 > x_2$, $r^{10}_1 > y_1$, and $y_2 > r^{10}_2$. Tversky and Kahneman find that changes in the reference point from r^1 to r^{10} shift preferences towards y. Following Tversky and Kahneman, we call this the *advantages/ disadvantages effect*.

Finally, compare r^2 and r^3, defined by $x_1 > r^3_1 > r^2_1$ and $r^3_2 = r^2_2 = x_2$. Herne finds that changes in the reference point from r^2 to r^3 shift preferences towards y. We shall call this the *relative closeness effect*. (The idea is that the difference between y_1 and x_1 appears greater when measured relative to r^3_1 than when measured relative to r^2_1.) Since the labelling of the goods as '1' and '2' is arbitrary, an equivalent statement of this effect is the following: changes in the reference point from r^9 to r^8 shift preferences towards x.

One striking feature of the experimental evidence is the speed with which reference points adjust to changes in holdings of goods. For example, in Knetch's experiment (described above), a few minutes of ownership of a chocolate bar or a coffee mug was enough to generate an additional degree of preference for it. Arguably, this phenomenon reveals a form of myopia: a person who has been given a mug regards a mug as preferable to a chocolate bar, even though, were he actually to take the chocolate in exchange, he would almost immediately prefer the chocolate to the mug. In fact, Loewenstein and Adler (1995) find some evidence that individuals fail to predict the changes in their preferences that result from changes in their holdings. This suggests that a theory of choice that

assumes reference-dependent preferences needs to treat reference points as endogenous.

2. Tversky and Kahneman's theory of reference-dependent preferences

Remarkably, all the regularities described in Section 1 are consistent with Tversky and Kahneman's theory of reference-dependent preferences; and all but one of them (the advantages/disadvantages effect) is positively predicted. We now outline that theory.

Throughout the rest of this paper, we consider an individual's preferences over all possible bundles of n goods (with $n \geq 2$), represented by the set $\mathbb{R}_+^n$. Typical elements in this set will be represented by r, s, x, y, z, where $r = (r_1, ..., r_n)$ and so on. A *preference relation* $\succsim$ is a binary relation on $\mathbb{R}_+^n$; $x \succsim y$ is read as 'x is weakly preferred to y'; the relations of strict preference ($\succ$) and of indifference ($\sim$) are defined from $\succsim$ in the usual way. In Hicksian consumer theory, an individual's preferences are described by a single preference relation. Instead, we use the more general concept of a *preference structure*, defined as a function from $\mathbb{R}_+^n$ to the set of all preference relations; to each *reference point* r in $\mathbb{R}_+^n$, a preference structure assigns a *reference-dependent preference* relation $\succsim_r$. The relation $\succsim_r$ describes the individual's preferences over consumption bundles when his reference point is r (or as we shall sometimes say, his preferences *viewed from* r). A preference structure is *reference-independent* if $\succsim_r$ is identical with $\succsim_s$ for all r, s. This special case corresponds with the treatment of preferences in Hicksian consumer theory.

It is convenient to use the notation x R y, x P y, x I y to denote $x \succsim_y y$, $x \succ_y y$, $x \sim_y y$ respectively. Thus, for example, x R y signifies that if the individual were endowed with y and viewed this as his reference point, he would be willing to exchange y for x. Notice that the relations R, P and I do not necessarily have the respective properties of $\succsim$, $\succ$ and $\sim$. For example, R is not necessarily complete (an individual might be willing neither to move from x to y nor to move from y to x); x I y does not entail y I x (an individual who is just willing to move from y to x might be unwilling to move from x to y). We use the notation x RR y to denote that there exists some sequence of bundles $z^1, ..., z^m$ in $\mathbb{R}_+^n$ such that z^1 R y, z^2 R z^1, ..., z^m R z^{m-1}, and x R z^m. Thus, x RR y signifies that if the individual is endowed with y, and if he always views his current endowment as his reference point, he is willing to engage in each of a series of exchanges leading from y to x.

The concept of a preference structure is the core idea in Tversky and Kahneman's theory.

In order to generate a useful theory, however, it is necessary to impose restrictions on the preference structure. One type of restriction applies to each $\succsim_r$, considered separately. Tversky and Kahneman assume:

(A1) For all r: $\succsim_r$ is complete (i.e. for all x, y: $x \succsim_r y$ or $y \succsim_r x$).

(A2) For all r: $\succsim_r$ is transitive.

(A3) For all r, x, y: if $x > y$ then $x \succ_r y$.

Significantly, convexity is *not* assumed.

The second type of restriction is concerned with how the reference-dependent preference ranking of any given (x, y) pair changes as the reference point changes. We shall say that a change in reference point from r to s *weakly favours* y relative to x if (i) $y \sim_r x$ implies $y \succsim_s x$ and (ii) $y \succ_r x$ implies $y \succ_s x$; it *strictly favours* y relative to x if $y \succsim_r x$ implies $y \succ_s x$. Notice that, given A1, the concept of 'favouring' is symmetrical in the following sense: the propositions 'a change in reference point from r to s weakly (respectively: strictly) favours y relative to x' and 'a change in reference point from s to r weakly (respectively: strictly) favours x relative to y' are equivalent to one another.

The distinctive features of Tversky and Kahneman's theory derive from three assumptions about the effects of changes in reference points. Consider any good $i \in \{1, ..., n\}$, any bundles x, y such that $y_i > x_i$, and any reference points r, s such that $s_i > r_i$ and $r_j = s_j$ for all $j \neq i$. Tversky and Kahneman assume:[1]

(A4) If $y_i \geq s_i$ and $x_i = r_i$, then the move from r to s strictly favours y relative to x.

(A5) If $x_i \geq s_i$, then the move from r to s weakly favours y relative to x.

(A6) If $r_i \geq y_i$, then the move from r to s weakly favours x relative to y.

The conditions formed by substituting 'strictly' for 'weakly' in A5 and A6 will be called the *strict versions* of A5 and A6; conversely, the condition formed by substituting 'weakly' for 'strictly' in A4 is the *weak version* of A4. Tversky and Kahneman describe A4 as a condition of 'loss aversion', A5 as a condition of 'diminishing sensitivity in gains', and A6 as a condition of 'diminishing sensitivity in losses'.

Before discussing the motivation of these conditions, we consider their implications for the

regularities in behaviour described in Section 1. It is easy to show the following results:

- The WTA/WTP effect, conservatism, the EG effect, the EL effect, and the inner asymmetric dominance effect are all implied by A4.
- The outer asymmetric dominance effect is implied by the conjunction of A4 and A5.
- The relative closeness effect is implied by the strict version of A5.
- The advantages/disadvantages effect is consistent with, but not implied by, A4–A6. A4 and A5 imply that a change in the reference point from r^3 to r^8 shifts preferences in favour of y (i.e. in the same direction as the advantages/disadvantages effect), but A6 implies that changes from r^1 to r^3 and from r^8 to r^{10} shift preferences in favour of x (i.e. in the opposite direction).

Notice that A6 plays no part in explaining the observed regularities.

To understand why Tversky and Kahneman impose A4–A6 on the preference structure, it is necessary to understand the main features of their earlier theory of choice under uncertainty, *prospect theory* (Kahneman and Tversky, 1979). Tversky and Kaheneman (1991, pp. 1039-1040) present the theory of reference-dependent preferences as an 'extension' of prospect theory. Prospect theory applies to choices over lotteries, all of whose consequences are gains or losses in a single dimension (which we may call *wealth*). In prospect theory, the objects of preference are probability distributions over changes in wealth. There is a *value function* v(.), unique up to multiplication by a positive constant with $v(0) = 0$ as a natural zero, which assigns a real-valued index v(w) to every increment or decrement of wealth w, measured relative to any given reference point. In addition to being continuous and strictly increasing, v(.) has the following properties:

(B1) For all w: $v(w) < -v(-w)$.

(B2) For all $w \geq 0$: v(.) is concave.

(B3) For all $w \leq 0$: v(.) is convex.

In prospect theory, B1 is called 'loss aversion', B2 'diminishing sensitivity for gains', and B3 'diminishing sensitivity for losses'. B1 is glossed as 'losses loom larger than corresponding gains'; it imparts a tendency, other things being equal, for individuals to be averse to prospects which involve risks of loss (measured relative to the reference point). B2 and B3 are glossed as 'the marginal value of both gains and losses decreases with their size'. B2 imparts a tendency towards risk

aversion in choices between prospects with positive outcomes; B3 imparts a tendency towards risk-loving behaviour in choices between prospects with negative outcomes. In the domain of choice under risk, there is evidence in support of all three of these tendencies.

Despite the common terminology, the connection between B1–B3 and A4–A6 is not immediately obvious: taken at face value, the two sets of conditions refer to different theoretical domains. The connection seems to be through the following special case of the theory of reference-dependent preferences, not presented explicitly by Tversky and Kahneman. Assume that, for each r, the relation $\succsim_r$ can be represented by the function u(x, r) where

$$u(x, r) = \sum_i v_i(x_i - r_i), \tag{1}$$

and where each $v_i(.)$ is a strictly increasing real-valued function, unique up to a multiplication by a positive constant, with $v_i(0) = 0$. We shall call this the *additive representation*. Given this representation, A4 is an implication of the additional assumption that each $v_i(.)$ satisfies B1 and B2. A5 and A6 are, respectively, implications of the assumptions that each $v_i(.)$ satisfies B2 and B3. Conversely, the conjunction of A4, A5, and A6 implies that each $v_i(.)$ satisfies B1, B2, and B3.

As we understand it, Tversky and Kahneman's approach is grounded in the additive representation.[2] However, they recognise that this representation is too restrictive, perhaps for the following reasons. First, as is well known in consumer theory, additive separability does not allow goods to be complements. Second, it does not allow the strength of status quo bias to vary with the degree of similarity between goods. For example, suppose goods 1, 2, 3 and 4 are respectively Cadbury chocolate, Nestlé chocolate, Pepsi, and Coke. Intuitively, it seems possible that a consumer might reveal status quo bias in exchanges between small quantities of chocolate (whatever the brand) and of cola drink (whatever the brand), while showing no such bias in exchanges between the two brands of chocolate, or between the two brands of drink. It is easy to show that such a pattern of reference-dependent preferences is inconsistent with the additive representation. Third, the assumption that preferences depend *only* on increments and decrements of consumption, independently of absolute levels, has unrealistic implications for behaviour in markets. Suppose that a person's reference point is given by his current holdings of goods and that, given the prices prevailing in the market, he would choose to engage in some trade – say, giving up some of his endowment of good i in return for some amount of good j. This trade constitutes a change in his

endowments. After his reference point has adjusted to this change, he will choose to make exactly the same trade again, giving up more of his endowment of good i. This will continue until his whole endowment of good i is exhausted.

Tversky and Kahneman's theory of preference structures is more general than the additive representation, and thus does not positively predict the effects described in the previous paragraph. However, it uses assumptions whose motivation derives from the additive special case. In particular, A6 is the manifestation in consumer theory of B3, an assumption that is used in prospect theory to explain risk-loving behaviour in choices over lotteries involving losses. As we have already noted, A6 seems to do no work in explaining observed regularities in consumer choice. But it lies behind two features of Tversky and Kahneman's theory of reference-dependent preferences which diverge sharply from conventional consumer theory: it permits non-convex preferences, and it permits cycles of choice of the form y P x, z P y, x P z.

First, consider convexity. It is easy to show that, given the additive representation, reference-dependent preferences are convex everywhere if and only if each $v_i(.)$ is concave everywhere. If each $v_i(.)$ is *convex* in its negative domain, as B3 requires, reference-dependent preferences are concave in the region of commodity space that is dominated by the reference point. This is not to say that, in the general theory proposed by Tversky and Kahneman, A6 *implies* that preferences are concave in this sense. Nevertheless, the additive special case illustrates how A6 imparts a tendency towards such concavity. Consider what patterns of behaviour would be observed if reference-dependent preferences were strictly concave in the region dominated by the reference point. Suppose that reference points are interpreted in terms of customary consumption. Then, if a consumer faced a budget constraint which passed just below the reference point – a case which would arise if a change in prices or income prevented him from consuming a bundle to which he had become accustomed – he would maintain his reference levels of consumption of all goods but one. We know of no evidence of such an effect.

Now consider choice cycles. The following example shows that Tversky and Kahneman's theory permits cycles. Let n = 2 and consider the three bundles x = (1, 3), y = (2, 2), z = (3, 1). Using the additive representation, let $v_1(-2) = -14$, $v_1(-1) = -13$, $v_1(1) = 12$, $v_1(2) = 13$, $v_2(-2) = -18$, $v_2(-1) = -10$, $v_2(1) = 9$, $v_2(2) = 17$. These indices are consistent with B1–B3. They imply u(y, x) = u(z, y) = 2 and u(x, z) = 3. Thus, y P x, z P y, and x P z. Intuitively, cycles are possible if

diminishing sensitivity for losses outweighs loss aversion, so that the sum of the utilities from a series of small gains can outweigh the disutility from a single loss, equal and opposite (in physical units) to the sum of those gains. This result suggests that A6 imparts some tendency towards choice cycles.

If (as the experimental evidence suggests) reference points adjust quickly to changes in the status quo, choice cycles would have destabilising effects on markets. For example, suppose that some individual has the preferences y P x, z P y, and x P z. Then, starting from a status quo position at x, he would have a positive desire to exchange x for y. But having adjusted to y as his new status quo, he would have a positive desire to exchange y for z; and so on. Thus, markets could induce an endless pass-the-parcel circulation of goods, limited only by the frictions generated by transaction costs. While there may be some particular cases in which consumers' choices show cyclical patterns over a relatively long time-scale (think of the effects of fashion), it does not seem credible to suppose that choice cycles are a normal property of individual behaviour, induced by the *general* properties of preference structures. In any event, we know of no evidence of choice cycles of the kind that might be induced by diminishing sensitivity.[3]

3. Organising the data on reference point effects

From the perspective of consumer theory, the set of restrictions that Tversky and Kahneman impose on preference structures seems less than ideal. On the one hand, one of their restrictions – A6, the condition of diminishing sensitivity for losses – is not needed to explain observed consumer behaviour. On the other hand, their theory permits non-convex preferences and choice cycles – phenomena that, as far as we know, have not been observed. We have shown that these two features of Tversky and Kahneman's theory are intimately related.

One possible response would be simply to drop A6. But since A5 and A6 have been given essentially the same psychological motivation in terms of diminishing sensitivity, that would be an ad hoc manoeuvre. Instead, we suggest the following general condition as an alternative to the conjunction of A4, A5 and A6:

> (A7) For all goods i, j $\in$ {1, ..., n}, for all bundles x, y such that $y_i > x_i$, $y_j < x_j$ and $x_k = y_k$ for all $k \neq i, j$, and for all reference points r, s such that $s_i > r_i$ and $r_k = s_k$ for all $k \neq i$: the move from r to s strictly favours y relative to x.

If reference-dependent indifference curves are smooth, A7 is equivalent to the condition that, at any given point in goods space, the marginal rate of substitution between good j and good i becomes more negative as r_i increases. Looking at Figure 1, it is easy to verify the truth of the following result:

RESULT 1: A7 implies the WTA/WTP effect, conservatism, the EG effect, the EL effect, the inner and outer asymmetric dominance effect, the relative closeness effect, and the advantages/ disadvantages effect.

We are reluctant to propose A7 as a fundamental property of the theory of reference-dependent preferences when we do not have any explanation of *why* reference points might affect preferences in this way, when the evidence for some of the behavioural regularities that A7 organises is provided by only one or two experiments, and when many significant predictions of A7 have never been tested. We prefer to regard this condition merely as a promising preliminary hypothesis. We shall use it as a screening device: in order to show that all the regularities described in Section 1 are consistent with any given theory, it is sufficient to show that A7 is consistent with that theory.

4. A new model of preference structures

Our strategy in this paper is to take Hicksian consumer theory as our template, and to amend it just as much as is necessary to accommodate the evidence on reference point effects. We use the concept of a preference structure and the associated notation, as defined in the second and third paragraphs of Section 2. But in place of Tversky and Kahneman's conditions A1–A6, we propose the following assumptions about the preference structure of any given individual:

(C1) *Completeness* For all r: $\succsim_r$ is complete.

(C2) *Transitivity* For all r: $\succsim_r$ is transitive.

(C3) *Increasingness* For all r, x, y: if $x > y$ then $x \succ_r y$.

(C4) *Strict convexity* For all r, x: $\{y \mid y \succsim_r x\}$ is strictly convex.

(C5) *Continuity for a given reference point* For all r, x: $\{y \mid y \succsim_r x\}$ and $\{z \mid x \succsim_r z\}$

are closed.

(C6) *Continuity for changes in reference points* For all x, y: $\{r \mid x \succsim_r y\}$ is closed.

(C7) *Weak acyclicity* For all x, y: $\neg(y \, RR \, x \wedge x \, P \, y)$.

C1–C5 require that each reference-dependent preference relation $\succsim_r$ satisfies all the properties that are normally assumed of preference relations in Hicksian consumer theory.

C6 and C7 impose the only restrictions that apply across reference points. The continuity condition C6 ensures that small changes in the reference point are associated with small changes in reference-dependent preferences. We suggest that this is a natural extension of the standard continuity condition for Hicksian preferences. C7 plays a major role in our theory, by restricting the ways in which reference-dependent preferences may vary with reference points. It is motivated partly by an analogy with revealed preference theory, and partly by a hypothesis about the psychological mechanisms underlying status quo effects.

First, the theoretical analogy. It is well known that the weak axiom of revealed preference is insufficient to provide a revealed-preference foundation for the utility function assumed in Hicksian consumer theory. In order to guarantee the existence of a utility function, it is necessary to impose the *strong* axiom of revealed preference. This axiom explicitly excludes the possibility of cycles of choice in which, for some x^1, ..., x^m, x^2 is chosen when x^1 is feasible, x^3 is chosen when x^2 is feasible, ... , and x^1 is chosen when x^m is feasible. This assumption is sometimes defended – for example, by Paul Samuelson (1950, p. 369) who first proposed it – on the grounds that a consumer who violated it would be vulnerable to a 'money pump', that is, to a sequence of trading opportunities, each of which is willingly undertaken, but which taken together lead to unambiguous loss. In our theory, C2 rules out cycles of preference *when the reference point is fixed,* but is silent about the patterns of exchanges that might be induced by endogenous reference points. C7 extends C2 by ruling out cycles of the following form: x^2 is chosen when x^1 is feasible *and is the reference point*, x^3 is chosen when x^2 is feasible *and is the reference point*, ... , and x^1 is chosen (and strictly preferred) when x^m is feasible *and is the reference point*. In this sense, C7 is a natural extension to preference structures of a fundamental assumption of conventional theory.

Notice that one implication of C7 is that reference-dependent preferences are not vulnerable to money pumps. One might expect that individuals who participate in markets would learn not to

have preferences that could be exploited by arbitrageurs. Money pump arguments are often used to challenge non-standard theories of preferences; because of C7, such arguments have no force against the theory we are presenting.

Now for the psychology. We suggest that the psychological mechanisms that induce reference point effects also tend to impede cycles. However these mechanisms are characterised, it seems clear that they induce some form of reluctance to move away from the status quo. In a cycle of the form we have just described, each step involves a decision to move away from the current status quo – a move which, other things being equal, is impeded by status quo bias. Notice also that a cycle with m = 2 (that is, x^2 is weakly preferred to x^1 when x^1 is the reference point, but x^1 is strictly preferred to x^2 when x^2 is the reference point) is the exact opposite of the regularity that is revealed in the WTA/WTP discrepancy and in conservatism.

We shall occasionally refer to two stronger variants of C7. Letting d(y, z) denote the Euclidian distance between y and z, these variant conditions are:

(C7*) *Strict acyclicity* For all distinct x, y: $\neg(y \; RR \; x \wedge x \; R \; y)$.

(C7**) *Limit acyclicity* For all x, and for all $\delta > 0$: there exists some $\varepsilon > 0$ such that, for all x^1, x^2, x^3 which satisfy $d(x^1, x) < \varepsilon$, $d(x^3, x) < \varepsilon$ and $d(x^2, x) > \delta$, $\neg(x^2 \; RR \; x^1 \wedge x^3 \; RR \; x^2)$.

C7* strengthens C7 by ruling out the possibility of a sequence of willing exchanges in which the individual starts with x, moves to y, and then returns to x. C7** is a still stronger condition; it rules out the possibility of a sequence of willing exchanges in which the individual starts with a bundle which is arbitrarily close to x, moves to y, and then returns to some other bundle arbitrarily close to x. Thus, while C7 merely *allows* preferences to vary with reference points in the direction that corresponds with status quo bias, C7* and C7** *postulate* that such bias exists.

5. A functional form for preference structures

If the model presented in Section 3 is to serve its intended purpose, it must be consistent with the evidence of reference point effects. Since A7 implies all the reference point effects we have described, it is sufficient to show that our model is consistent with A7. That this is the case is

established by the following result:

RESULT 2 Preference structures exist which satisfy C1–C7 and A7.

We shall now present a class of preference structure which constitutes a proof of Result 2.[4] This example serves also to illustrate some general features of our model, and to show that reference-dependence can be incorporated into tractable functional forms. Consider the following *CES preference structure*, which generalises a familiar functional form for the representation of preferences:

$$u(x, r) = A(r) [\sum_i \gamma_i r_i^{\rho - \beta} x_i^{\beta}]^{1/\beta} \qquad (2)$$

with $\sum_i \gamma_i = 1$ and $1 > \rho \geq \beta > -\infty$; A(.) is any continuous function such that $A(r) > 0$ for all r.

First, consider the implications of (2) for reference-dependent preferences in relation to any *fixed* reference point r. In this context, A(r) is merely an arbitrary normalisation. Then u(x, r) is a reference-dependent CES (constant elasticity of substitution) utility function; the elasticity of substitution is $1/(1 - \beta)$. It is well known that preferences represented by such a function satisfy C1–C5. Thus, (2) implies that the preference structure is a family of reference-dependent preference relations $\succsim_r$, each of which satisfies the restrictions imposed by our model.

Now consider how reference-dependent preferences vary with changes in reference points. It is clear from the form of (2) that C6 is satisfied. In addition, the differentiability properties of (2) imply that indifference surfaces are smooth everywhere. This smoothness property can be used to define a relation of long-run preference. For any reference point r, the set of bundles $I(r) = \{x \mid x \, I \, r\}$ is an indifference surface of the reference-dependent preference ordering $\succsim_r$. The local properties of this indifference surface at r represent the consumer's preferences at r, viewed from r itself. A *long-run indifference surface* is a surface in $\mathbb{R}^n_+$ such that, at each point r on this surface, the local properties of the surface are the same as those of the corresponding I(r). Thus, long-run indifference surfaces are constructed from local properties of I(r) surfaces in much the same way that, in revealed preference theory, indifference surfaces are constructed from a consumer's market choices. If a family of such surfaces can be constructed, it induces a *long-run preference* relation $\succsim_L$.

But can such surfaces always be constructed? This is the *integrability problem* explained

by Samuelson (1950). In our general model, C7 (whose role is analogous with that of the strong axiom in revealed preference theory) guarantees the existence of long-run indifference surfaces if reference-dependent indifference surfaces are smooth. But since we have not yet shown that (2) satisfies C7, we proceed by demonstration.

In fact, the long-run preferences induced by (2) are represented by the long-run utility function

$$v(x) = [\sum_i \gamma_i x_i^{\rho}]^{1/\rho}. \qquad (3)$$

This is a CES function whose elasticity of substitution is $1/(1-\rho)$. It is straightforward to verify that, at any reference point r, the long-run marginal rates of substitution (MRSs) implied by (3) are equal to the reference-dependent MRSs implied by (2), and hence that (3) does indeed represent long-run preferences. Since (3) is a CES function, the long-run preference relation $\succsim_L$ is complete, transitive, increasing, strictly convex, and continuous. It is convenient to use the normalisation

$$A(r) = [\sum_i \gamma_i r_i^{\rho}]^{1/\rho - 1/\beta}, \qquad (4)$$

so that the substitution of r = x into (4) and (2) yields (3).

The condition $\rho \geq \beta$ ensures that reference-dependent indifference surfaces are 'at least as convex as' long-run indifference surfaces; if $\rho > \beta$, the former surfaces are 'more convex' than the latter. Figure 2 shows typical surfaces for the case of two goods with $\rho > \beta$ (r and s are different reference points, I(r) and I(s) are the corresponding reference-dependent indifference surfaces, and I_L is a long-run indifference surface). Notice that for any distinct x, y, if x is on or above I(y), x must also be on or above the long-run indifference surface which passes through y. Thus, x R y implies x $\succsim_L$ y, and x P y implies x $\succ_L$ y. Since $\succsim_L$ is transitive, the conjunction of y RR x and x P y implies a contradiction. Thus, C7 is satisfied. If $\rho > \beta$, C7* is also satisfied.[5]

We now show that if $\rho > \beta$, A7 is satisfied. Let $M_{kj}(x, r)$ denote the marginal rate of substitution (MRS) between goods k and j, evaluated at point x, viewed from the reference point r. For all k, j, x, r:

$$M_{kj}(x, r) = -(\gamma_j/\gamma_k)(r_j/r_k)^{\rho-\beta}(x_j/x_k)^{\beta-1}. \qquad (5)$$

Thus if $\rho > \beta$, the MRS between any two goods k and j at any given point x becomes more negative as r_j increases, everything else being held constant. This is A7.

6. Making reference points endogenous: preliminary remarks

We now consider some of the implications of our model for consumer behaviour, treating reference points as endogenous. If we are to do this, we must choose whether to interpret a consumer's reference point as his endowment or as his customary consumption. We suggest that which interpretation is more appropriate depends on the situation being modelled. Consider two paradigm cases.

The first case is that of a consumer who, in each of many periods, chooses between alternative bundles *for consumption in that period*, facing the same budget constraint in every period. In any given period, his endowment is the amount of money he has to spend in that period, while his customary consumption is some function of the bundles he chose to buy in preceding periods. When we speak of making his reference point endogenous, we mean to take account of how his preferences might change *from consumption period to consumption period*. In this case, it seems most natural to interpret reference points as customary consumption.

The second case is that of an individual who, over a relatively short period of time, faces a series of opportunities to buy or sell *stocks* of goods, the services of which will be consumed later. This case is typical of most of the experimental environments in which status quo effects have been observed. For an example from outside the laboratory, consider a person buying and selling shares. Here, when we speak of making the individual's reference point endogenous, we mean to take account of how his preferences might change *over the course of a trading process*. In this case, it seems most natural to interpret reference points as endowments.

In this paper, we focus on models in which reference points are identified with endowments; these models are to be understood as representing economic situations similar to our second paradigm case. Our main reason for not giving equal attention to the 'customary consumption' interpretation of reference points is that, under this interpretation, endogenous reference-dependence is formally equivalent to *habit formation*, on which a substantial theoretical literature already exists. In the remainder of this Section, we briefly explain this formal analogy, note some of its implications, and explain why the analogy does not extend to the 'current endowment' interpretation of reference points.

In the habit formation literature, individuals' preferences are assumed to adapt to previous consumption experiences. This adaptive process is typically interpreted either as a model of addiction (for example: consumption of nicotine in one period induces nicotine dependence in later periods) or as a model of the accumulation of 'consumption capital' (for example: consumption of music in one period induces greater ability to enjoy music in later periods). The psychological explanation of status quo effects is quite different; but if reference points are construed as customary consumption, a model with endogenous reference points has the same formal structure as a model of habit formation. The implications of such models for consumer demand are examined in a literature initiated by Robert Pollak (1970) and Christian von Weizsäcker (1971).

The general equilibrium implications of reference-dependence, interpreted as customary consumption, can be explored by adapting the formal analysis developed in another literature, that of *general equilibrium with nontransitive preferences*. This literature dispenses with the conventional assumption that preferences are complete. For each individual, a strict preference relation ™, which need not be transitive, is postulated; it is assumed that, for each point r in goods space, the set of bundles strictly preferred to r is non-empty and convex. An economic agent is deemed to be optimising if he chooses a bundle x* such that no strictly preferred bundle is feasible. David Gale and Andrew Mas-Collel (1975) and Wayne Shafer and Hugo Sonnenschein (1975) prove, for every fixed profile of individuals' endowments, the existence of a price vector such that all markets can clear with individuals optimising in this sense. Interpreting ™ as the reference-dependent relation P, we can use this result to establish the existence of long-run general equilibrium in a reference-dependent model in which *individuals' endowments are constant over time* but their reference points adapt to consumption experience.

However, if reference points are interpreted as current endowments, a different conceptualisation of equilibrium is needed. On this latter interpretation, endowments change as trade proceeds. Trade has to be modelled as a sequential process; at any given time, an individual takes his current holdings of goods as his reference point in choosing what trades to make; but these trades then induce changes his holdings, and thus induce changes in his reference point. We now investigate the properties of such processes.

7. Individual choice when reference points are endogenous

From now on, we interpret reference points as current endowments. We shall assume that an individual's decisions in any trading period are determined by his reference-dependent preferences, his endowment at the start of the period acting as the reference point. We shall also assume that, in forming trading plans for any period, individuals do not take account of the effects that their trades will have on their reference points, and hence on their future preferences.

In this section, we consider the trading behaviour of an individual who, over a series of periods, faces a fixed set of trading opportunities. We first define a *feasible set* as a nonempty, closed, bounded and convex set $S \subset \mathbb{R}_+^n$, satisfying the restriction that, for all $x, y \in \mathbb{R}_+^n$: $(x \in S \wedge x > y) \Rightarrow y \in S$. For a given S, the corresponding *exchange set* X is the set of nondominated elements of S, i.e. $X = \{y \mid y \in S \wedge (\neg\exists z)(z \in S \wedge z > y)\}$. We interpret the exchange set as a set of bundles, every one of which is in every period exchangeable for any other. For example, suppose the individual enters a market with certain initial holdings of goods y, and faces a fixed vector of prices $p = (p_1, ..., p_n)$ at which he can buy and sell goods 1, ..., n. Then $S = \{x \mid \sum_i p_i (x_i - y_i) \leq 0\}$ is a feasible set in our sense, and $X = \{x \mid \sum_i p_i (x_i - y_i) = 0\}$ is an exchange set.

We now fix a particular feasible set S, and thereby an exchange set X. We also fix the individual's preference structure P, and stipulate that this satisfies C1–C7. We define a *trading sequence* as a sequence $<x(0), x(1), ...>$ such that each x(t) is a member of X. We interpret each x(t) as the individual's endowment in period t; the difference between x(t) and x(t + 1) represents the exchanges carried out in period t.

We define a *reflexive optimum* as a bundle x* that is optimal in X, viewed from itself. Notice that, because reference-dependent preferences are increasing and strictly convex, any such x* is *uniquely* optimal, both in X and in S. A reflexive optimum can be interpreted as an equilibrium state of the individual's holdings: were the individual to be endowed with such a bundle in any period, he would not wish to engage in any further exchanges. That a reflexive optimum exists follows from a simple application of Brouwer's fixed-point theorem. (We can define a mapping g(.) from S to itself, such that each g(x) is the uniquely optimal bundle in S, viewed from x. Clearly, each g(x) is an element of X. Because of the continuity properties of P, g(.) is a continuous function. Thus, it has a fixed point. Any such point is a reflexive optimum.) It is natural to ask whether,

starting from an arbitrary x(0), a trading sequence will converge to a reflexive optimum.

Clearly, if the answer to this question is 'Yes', this must be by virtue of some degree of rationality on the part of the individual. Even for an individual with a reference-independent preference structure – that is, for an individual as modelled in conventional consumer theory – his ability to reach an optimum by means of trade depends on his acting on certain principles of rationality. In this paper, our concern is with the *particular* problems that reference-dependence introduces into consumer theory. Accordingly, our strategy is to assume that the individual acts on certain minimal principles of rationality, sufficient to lead him to an optimum if he has a reference-independent preference structure. We then investigate whether the same principles of rationality lead to a reflexive optimum when preferences are reference-dependent.

We begin by imposing the following natural restriction on trading sequences:

(D1) *Local improvement* For all $t > 0$: (i) $x(t)\ R\ x(t-1)$ and (ii) if $x(t-1)$ is not a reflexive optimum, then $x(t)\ P\ x(t-1)$.

D1 requires that, in every period, the individual takes *some* advantage of available opportunities for gain, as viewed from his current reference point.[6]

However, even with reference-independent preferences, D1 is not sufficient to guarantee convergence to optimality. There are perverse cases of trading sequences which, while in every period capturing some non-zero part of the current opportunities for gain, do not converge to optimality. Intuitively, we need a condition which requires that in each period the individual captures a *non-trivial* part of the current opportunities for gain.

As a means of formulating such a restriction, we define a measure of gains from trade. For all sets $Z \subseteq \mathbb{R}_+^n$, we use L(Z) to denote the Lebesgue measure of Z. Taking S, X and P as given, we define a set-valued function $\Phi(.,.)$ as follows: For all x, z in S, $\Phi(z, x) = \{y \mid y\ R\ x \wedge z > y\}$. To interpret this set, suppose that in some period the individual is endowed with x and is considering exchanging this endowment for z. $\Phi(x, z)$ contains all those bundles that are weakly preferred to x, viewed from x, but that are unambiguously inferior to z. (Notice that, because each of these bundles is dominated by z, and because z is an element of S, each bundle is itself in S, i.e. $\Phi(x, z) \subseteq S$.) Thus, if $z\ R\ x$ is true, $L(\Phi[z, x])$ can be interpreted as an index, in terms of preferences viewed from x, of how far z lies above the indifference surface that passes through x. In other words, it can be

interpreted as an index of the *gains from trade* that the individual realises in moving from x to z, evaluated with respect to his reference point in the relevant trading period. If z R x is false, $L(\Phi[z, x]) = 0$.

Our second restriction on trading sequences is:

(D2) *Non-trivial improvement* If $L(\Phi[x(t + 1), x(t)]) \rightarrow 0$ as $t \rightarrow \infty$, then also $\max_{z \in X} [L(\Phi[z, x(t)])] \rightarrow 0$ as $t \rightarrow \infty$.

D2 requires that *actual* gains from trade in each period do not become vanishingly small unless this is also true of *potential* gains from trade.

The following convergence result can be proved (the proofs of this and subsequent results are given in the Appendix):

RESULT 3 For all preference structures $\mathcal{P}$ which satisfy C1–C7, and for all x(0) $\in$ X: every trading sequence that satisfies D1 and D2 converges to the set of reflexive optima.

Notice that Result 3 does not establish that every trading sequence converges to *a* reflexive optimum. Of course, it is an immediate corollary of Result 3 that if there is a *unique* reflexive optimum, every trading sequence that satisfies D1 and D2 must converge to it. For reflexive equilibrium to be unique, it is sufficient that the preference structure induces a strictly convex long-run preference relation (as the CES structure does).

Nevertheless, there are examples of preference structures and exchange sets such that a trading sequence that satisfies D1 and D2 can converge towards a cyclical path that is always close to *some* reflexive optimum yet does not converge to any *particular* one. This is still possible if strict rather than weak acyclicity (i.e. C7* rather than C7) is assumed. However, if the stronger condition of *limit* acyclicity (C7**) is substituted for C7, or if the preference structure is reference-independent, convergence to an optimum *is* guaranteed:

RESULT 4 For all preference structures $\mathcal{P}$ which satisfy C1–C5 and which either *satisfy C6 and C7*** or *are reference-independent, and for all x(0) $\in$ S: every trading sequence that satisfies D1 and D2 converges to some reflexive optimum.*

8. Exchange when reference points are endogenous

So far, we have considered the trading behaviour of a single individual, facing an exogenously given opportunity set. We now extend this analysis to investigate how trade between individuals is affected by endogenous reference points.

We consider an exchange economy with n goods and N individuals (with $N \geq 2$). A typical bundle of goods possessed by individual j is denoted by $x^j = (x^j_1, ..., x^j_n)$. For each person j, preferences over such bundles are described by a preference structure P^j; the relations R, P, I and RR relevant to person j are written as R^j, P^j, I^j and RR^j. Notice that in defining each individual's preferences over *his own* bundles (rather than over the allocation of goods to all individuals) we are implicitly assuming that there are no externalities. An *allocation* is an Nn-dimensional vector $(x^1_1, ..., x^1_n; x^2_1, ..., x^2_n; ... ; x^N_1, ..., x^N_n)$; **x** denotes a typical allocation. Feasibility for the economy is defined in terms of a *resource constraint* $q = (q_1, ..., q_n)$ such that, for each good i, q_i is strictly positive and finite. Taking some resource constraint q as given, an allocation **x** is *feasible* if $\sum_j x^j_i \leq q_i$ for all goods i; the set of feasible allocations is A. The set of *exchangeable* allocations Y is the set of non-dominated elements of A. Thus, an allocation **x** is exchangeable if and only if $\sum_j x^j_i = q_i$ for all i. The underlying idea is that each of the exchangeable allocations can be transformed into any other by means of exchanges of goods between individuals.

We now fix a particular resource constraint q, and hence fix A and Y. We also fix the preference structures $P^1, ..., P^N$ of the N individuals, and stipulate that each P^j satisfies C1–C7. For all allocations **x** and **y**, **y** is a *Pareto-improving move from* **x** (denoted **x** P* **y**) if $y^j R^j x^j$ is true for all persons j and if $y^j P^j x^j$ is true for at least one j. If $y^j I^j x^j$ is true for all j, **y** is a *Pareto-indifferent move from* **x** (denoted **x** I* **y**). A feasible allocation **x** is a *reflexive Pareto optimum* if there exists no feasible **y** such that **y** P* **x**. Notice that if individuals are endowed with an allocation which is a reflexive Pareto optimum, no further voluntary trade can occur. In this sense, a reflexive Pareto optimum in an exchange economy is analogous with a reflexive optimum in the context of individual choice: it provides a natural concept of efficiency in an exchange economy.

We define a *collective trading sequence* as a sequence $<\mathbf{x}(0), \mathbf{x}(1), ... >$ such that each **x**(t) is an exchangeable allocation. We interpret each **x**(t) as a specification of individuals' holdings at the start of period t; the difference between **x**(t + 1) and **x**(t) represents exchanges carried out between individuals in period t. We do not propose any explicit model of the trading mechanism

which determines each $\mathbf{x}(t + 1)$, given its preceding $\mathbf{x}(t)$. Thus, for example, prices and excess demands do not appear explicitly in our model. Our approach is more general: we merely specify two minimal restrictions on collective trading sequences – restrictions that are analogous with D1 and D2, and that are consistent with a wide range of alternative models of the trading mechanism. We investigate whether these restrictions are sufficient to ensure convergence to reflexive Pareto optimality.

Our first restrictions is:

(E1) *Local improvement* For all $t > 0$: (i) either $\mathbf{x}(t)$ P* $\mathbf{x}(t - 1)$ or $\mathbf{x}(t)$ I* $\mathbf{x}(t - 1)$, *and* (ii) if $\mathbf{x}(t - 1)$ is not a reflexive Pareto optimum, then $\mathbf{x}(t)$ P* $\mathbf{x}(t - 1)$.

The first part of E1 requires that the overall effect of the exchanges carried out in any period is weakly beneficial to all individuals, viewed from their current reference points. Given our fundamental assumption of myopia, and given our implicit assumption of no externalities, this is an essential property of any trading mechanism in which individual participation is voluntary. The second part of E1 requires that, if potential gains from trade exist in period t, some part of those gains are realised in that period.

Our second restriction requires that in each period a *non-trivial* part of current potential gains from trade are realised. First, we extend $\Phi(.,.)$ to allocations as follows. For all $\mathbf{z}$, $\mathbf{x}$ in Y, we define $\Phi(\mathbf{z}, \mathbf{x}) = \{\mathbf{y} \mid \mathbf{y}\ P^*\ \mathbf{x} \wedge \mathbf{z} > \mathbf{y}\}$. Thus, $\Phi(\mathbf{z}, \mathbf{x})$ contains all those allocations $\mathbf{y}$ that are Pareto-improving moves from $\mathbf{x}$ but that are unambiguously inferior to $\mathbf{z}$ – in the sense that $\mathbf{z}$ gives every individual at least as much of each good as $\mathbf{y}$ does, and gives at least one individual more of at least one good. (Notice that, because each allocation in $\Phi(\mathbf{x}, \mathbf{z})$ is dominated by $\mathbf{z}$, and because $\mathbf{z}$ is an element of A, each allocation is itself in A, i.e. $\Phi(\mathbf{x}, \mathbf{z}) \subseteq$ A.) Thus, if $\mathbf{z}$ P* $\mathbf{x}$ is true, $L(\Phi[\mathbf{z}, \mathbf{x}])$ can be interpreted as an index of the gains from trade that the N individuals together achieve in moving from $\mathbf{x}$ to $\mathbf{z}$, evaluated with respect to their reference points in the relevant trading period. If $\mathbf{z}$ P* $\mathbf{x}$ is false, $L(\Phi[\mathbf{z}, \mathbf{x}]) = 0$. Hence our second restriction:

(E2) *Non-trivial improvement* If $L(\Phi[\mathbf{x}(t + 1), \mathbf{x}(t)]) \to 0$ as $t \to \infty$, then also $\max_{\mathbf{z} \in Y} L(\Phi[\mathbf{z}, \mathbf{x}(t)]) \to 0$ as $t \to \infty$.

E2 requires that actual gains from trade in each period do not become vanishingly small unless this is also true of potential gains from trade.

The following convergence results, analogous with Results 3 and 4 for individual trading behaviour, can be proved:

RESULT 5 For all profiles of preference structures $P^1, ..., P^N$ *which satisfy C1–C7, and for all* $\mathbf{x}(0) \in Y$*: every collective trading sequence which satisfies E1 and E2 converges to the set of reflexive Pareto optima.*

RESULT 6 For all profiles of preference structures $P^1, ..., P^N$*, such that each* P^i *satisfies C1–C5 and* either *satisfies C6 and C7*** or *is reference-independent, and for all* $\mathbf{x}(0) \in Y$*: every collective trading sequence which satisfies E1 and E2 converges to some reflexive Pareto optimum.*

9. Conclusions

Our objective has been to develop a theory of consumer choice that is compatible with what is known about reference point effects, while conserving as much as possible of the generality and tractability of conventional consumer theory. Our strategy has been to postulate that preferences are conditional on reference points, and that, conditional on any given reference point, preferences have the same properties of completeness, transitivity, increasingness and convexity as are assumed in conventional theory. We have proposed two conditions that link preferences conditional on different reference points: a condition of continuity, and a condition which rules out cycles of choice. We have shown that preference structures with these properties can accommodate observed reference point effects. Under the special assumption that indifference surfaces are smooth, these effects occur when reference-dependent indifferent surfaces are more convex than 'long-run' indifference surfaces, that is, when substitution effects are weaker in the short run than in the long.

Given this theoretical framework, equilibrium for an individual who faces a given set of trading opportunities is naturally understood as a 'reflexive optimum': a bundle of goods that is optimal when it is 'viewed from itself' (that is, when this bundle is itself the reference point). Efficiency for an exchange economy is naturally understood as an allocation that is a 'reflexive Pareto optimum', that is, such that, when that allocation serves as the reference point for all

individuals, there are no unexploited gains from trade. We have shown that our assumptions about preference structures, combined with certain weak assumptions about the properties of trading mechanisms, are sufficient to ensure that sequences of trading decisions converge to states of reflexive optimality.

More generally, we hope that our work will persuade more economists to investigate the explanatory power of the hypothesis that preferences are reference-dependent. Many economists are reluctant to use non-standard assumptions about preferences, fearing that to do so would mean giving up the clarity, rigour and tractability of existing theory. Our results suggest that it is possible to model reference-dependence by means of relatively small modifications to the conventional theory of preferences, and that these modifications need not frustrate economists' attempts to build general theories of the workings of markets.

Appendix: Proofs of theorems

Proof of Result 3

Consider any preference structure P which satisfies C1–C7, any feasible set S and its corresponding exchange set X, and any trading sequence Q = <x(0), x(1), ... > which satisfies D1 and D2. For all $x \in X$, we define $C(x) = \{y \mid y \in S \wedge (\exists x^1, ..., x^m \in S: x^1 R x, x^2 R x^1, ... , x^m R x^{m-1}, y R x^m)\}$. Consider the function f(t) = L(C[x(t)]). The value of f(t) is bounded above by L(S), which is positive and finite, and below by zero. From the assumptions that reference-dependent preferences are strictly convex (C4) and that S is convex, it follows that for all x, x is a reflexive optimum if and only if C(x) = {x}. Thus, for all t, x(t) is a reflexive optimum if and only if f(t) = 0. Consider any t such that x(t) is not a reflexive optimum. By D1, x(t + 1) P x(t). Because reference-dependent preferences satisfy increasingness (C3) and weak acyclicity (C7), $\Phi(x[t + 1], x[t]) \cap C(x[t + 1]) = \varnothing$ for all t. But $\Phi(x[t + 1], x[t]) \subseteq C(x[t])$. Thus $f(t) - f(t + 1) \geq L(\Phi[x(t + 1), x(t)]) \geq 0$. So as t increases, f(t) does not increase. Thus, *either* (i) there exists some t' such that f(t) = 0 for all $t \geq t'$, *or* (ii) $f(t) \rightarrow 0$ as $t \rightarrow \infty$, *or* (iii) f(t) tends to some strictly positive limit as $t \rightarrow \infty$. If (i) is true, x(t) is a reflexive optimum for all $t \geq t'$. Suppose (ii) or (iii) is true. Then as $t \rightarrow \infty$, $f(t) - f(t + 1) \rightarrow 0$, which implies $L(\Phi[x(t + 1), x(t)]) \rightarrow 0$. Then by D2, $\max_{z \in X} [L(\Phi[z, x(t)])] \rightarrow 0$. Because P satisfies C1–C6, $\max_{z \in X} [L(\Phi[z, x])]$ is a continuous function of x,

whose value is zero if and only if x is a reflexive optimum. Thus $\max_{z \in X} [L(\Phi[z, x(t)])] \rightarrow 0$ as $t \rightarrow \infty$ if and only if Q converges to the set of reflexive optima. ?

Proof of Result 4

Consider any preference structure P which satisfies C1–C5 and which *either* satisfies C6 and C7** (case 1) *or* is reference-independent (case 2). Notice that in case 2, C6 and C7 are satisfied trivially. Now consider any feasible set S and its corresponding exchange set X, and any individual trading sequence Q = <x(0), x(1), ... > which satisfies D1 and D2. By Result 3, Q converges to the set of reflexive optima. If case 2 applies, there is a unique reflexive optimum (because of strict convexity), to which Q must converge. Suppose that case 1 applies and that Q does not converge to any single point. Since Q is bounded, by the Bolzano-Weierstrass Theorem it has a convergent subsequence Q′. Let the limit of Q′ be z. For all x, z, let d(x, z) denote the Euclidean distance between x and z. By supposition, Q does not converge to z. Thus, there exists $\delta > 0$ such that, for all $t > 0$, there exists $t' > t$ such that $d(x[t'], z) > \delta$. Fix any such δ. Now consider any $\varepsilon > 0$. Because Q′ converges to z, there must be some $t^1 > 0$ such that $x(t^1)$ is in Q′ and $d(x[t^1], z) < \varepsilon$. Because Q does not converge to z, there must be some $t^2 > t^1$ such that $d(x[t^2], z) > \delta$. Because Q′ converges to z, there must be some $t^3 > t^2$ such that $x(t^3)$ is in Q′ and $d(x[t^3], z) < \varepsilon$. By D2, $x(t^2)$ RR $x(t^1)$ and $x(t^3)$ RR $x(t^2)$. But the conclusion that such t^1, t^2, t^3 can be found for *all* $\varepsilon > 0$ is inconsistent with C7**. Thus the original supposition is false; Q converges to a point. Since Q converges to the set of reflexive optima, its limit is a reflexive optimum. ?

Proof of Result 5

Consider any profile of preference structures P^1, ..., P^N, all of which satisfy C1–C7, any resource constraint q and its corresponding sets A and Y, and any collective trading sequence Q = <$\mathbf{x}$(0), $\mathbf{x}$(1), ... > which satisfies E1 and E2. For all $\mathbf{x} \in Y$, we define $C(\mathbf{x}) = \{\mathbf{y} \mid \exists \mathbf{x}^1, ..., \mathbf{x}^m \in S: \mathbf{x}^1 P^* \mathbf{x}, \mathbf{x}^2 P^* \mathbf{x}^1, ... , \mathbf{x}^m P^* \mathbf{x}^{m-1}, \mathbf{y} P^* \mathbf{x}^m\}$. The continuation of the proof is identical with the corresponding continuation of the proof of Result 3, except that $\mathbf{x}$(t), $\mathbf{z}$, A, Y, ‘reflexive Pareto optimum’ and ‘E2’ are substituted for x(t), z, S, X, ‘reflexive optimum’ and ‘D2’ respectively. ?

Proof of Result 6

Consider any profile of preference structures P^1, ..., P^N, such that each P^i satisfies C1–C5 and *either* satisfies C6 and C7** *or* is reference-independent. Let $J \subseteq \{1, ..., N\}$ be the set of individuals whose preference structures satisfy C6 and C7**. Let $K \subseteq \{1, ..., N\}$ be the set of individuals whose preference structures are reference-independent. Notice that $J \cap K = \varnothing$. Notice also that, for each $k \in K$, R^k is a complete, transitive, increasing, strictly convex and continuous preference relation. Finally, notice that, for all individuals i, if P^i is reference-independent, it necessarily satisfies C6 and C7. Now consider any resource constraint q and its corresponding set of feasible allocations A and set of exchangeable allocations Y, and consider any collective trading sequence $Q = <\mathbf{x}(0), \mathbf{x}(1), ... >$ that satisfies E1 and E2. By Result 5, Q converges to the set of reflexive Pareto optima.

Suppose that Q does not converge to a single point. Since Q is bounded, it has a convergent subsequence Q′. Let the limit of Q′ be **z**. By supposition, Q does not converge to **z**. Thus, there exists $\delta > 0$ such that, for all t, there exists $t' > t$ such that $d(x[t'], \mathbf{z}) > \delta$. Fix any such δ. Now consider any $\varepsilon > 0$. Let $t^1(\varepsilon)$ be the lowest value of t such that $t \geq 0$, $\mathbf{x}(t)$ is in Q′, and $d(\mathbf{x}[t], \mathbf{z}) < \varepsilon$; because Q′ converges to **z**, such a value exists. Let $t^2(\varepsilon)$ be the lowest value of t such that $t > t^1(\varepsilon)$ and $d(\mathbf{x}[t], \mathbf{z}) > \delta$. Let $t^3(\varepsilon)$ be the lowest value of t such that $t > t^2(\varepsilon)$, $\mathbf{x}(t)$ is in Q′, and $d(\mathbf{x}[t], \mathbf{z}) < \varepsilon$; because Q′ converges to **z**, such a value exists. Notice that, as $\varepsilon \rightarrow 0$, $\mathbf{x}(t^1[\varepsilon])$ and $\mathbf{x}(t^3[\varepsilon])$ both converge to **z**. Hence, for all individuals i, $d(x^i[t^1(\varepsilon)], z^i) \rightarrow 0$ and $d(x^i[t^3(\varepsilon)], z^i) \rightarrow 0$ as $\varepsilon \rightarrow 0$. Because of E1, $x^i[t^3(\varepsilon)]\ RR^i\ x^i[t^2(\varepsilon)]$ and $x^i[t^2(\varepsilon)]\ RR^i\ x^i[t^1(\varepsilon)]$ are true for all i and for all $\varepsilon > 0$. Thus, and because P^i satisfies C7**, $d(x^j[t^2(\varepsilon)], z^j) \rightarrow 0$ as $\varepsilon \rightarrow 0$ for all individuals $j \in J$. So, in the limit as $\varepsilon \rightarrow 0$, the allocations $\mathbf{x}(t^1[\varepsilon])$, $\mathbf{x}(t^2[\varepsilon])$ and $\mathbf{x}(t^3[\varepsilon])$ differ only in respect of the bundles held by members of K.

Now consider any person $k \in K$. Because of E1, $x^k[t^2(\varepsilon)]\ R^k\ x^k[t^1(\varepsilon)]$ is true for all $\varepsilon > 0$. Because Q′ is a subsequence of Q which converges to **z**, because of continuity, and because of E1, $z^k\ R^k\ x^k[t^2(\varepsilon)]$ is also true for all $\varepsilon > 0$. Thus, $x^k[t^2(\varepsilon)]$ lies on or above the (reference-independent) indifference surface for person k that passes through $x^k[t^1(\varepsilon)]$, and on or below the indifference surface that passes through z^k. Since $x^k[t^1(\varepsilon)] \rightarrow z^k$ as $\varepsilon \rightarrow 0$, the distance between

$x^k[t^2(\varepsilon)]$ and the indifference surface through z^k tends to zero as $\varepsilon \to 0$. Notice that this result holds for all $k \in K$. Because the preferences of all members of K are continuous and strictly convex, and because (holding constant the bundles held by members of J) **z** is Pareto-optimal for the members of K, this result implies that $d(x^k[t^2(\varepsilon)], z^k) \to 0$ as $\varepsilon \to 0$ for all individuals $k \in K$.

Combining the conclusions of the preceding two paragraphs, $d(\mathbf{x}[t^2(\varepsilon)], \mathbf{z}) \to 0$ as $\varepsilon \to 0$. But $d(\mathbf{x}[t^2(\varepsilon)], \mathbf{z}) > \delta$ for all ε, a contradiction. Therefore the supposition that Q does not converge to a single point is false. ?

References

Bateman, Ian, Alistair Munro, Bruce Rhodes, Chris Starmer and Robert Sugden (1997). A test of the theory of reference-dependent preferences. *Quarterly Journal of Economics* 112: 479-505.

Bishop, Richard C. and Thomas A. Heberlein (1979). Measuring values of extramarket goods: are indirect measures biased? *American Journal of Agricultural Economics* 61: 926-930.

Gale, David and Andrew Mas-Colell (1975). An equilibrium existence theorem for a general model without ordered preferences. *Journal of Mathematical Economics* 2: 9-15.

Herne, Kaisa (1998). Testing the reference-dependent model: an experiment on asymmetrically dominated reference points. *Journal of Behavioral Decision Making* 11: 181-192.

Kahneman, Daniel, Jack L. Knetsch and Richard H. Thaler (1990). Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy* 98: 1325-1348.

Kahneman, Daniel and Tversky, Amos (1979). Prospect theory: an analysis of decision under risk. *Econometrica* 47: 263-291.

Knetsch, Jack L. (1989). The endowment effect and evidence of nonreversible indifference curves. *American Economic Review* 79: 1277-1284.

Knetsch, Jack L. and J.A. Sinden (1984). Willingness to pay and compensation demanded: experimental evidence of an unexpected disparity in measures of value. *Quarterly Journal of Economics* 99: 507-521.

Loewenstein, George and Daniel Adler (1995). A bias in the prediction of tastes. *Economic Journal* 105: 929-937.

Pollak, Robert A. (1970). Habit formation and dynamic demand functions. *Journal of Political Economy* 78: 745-163.

Rowe, Robert D., Ralph C. d'Arge and David S. Brookshire (1980). An experiment on the economic value of visibility. *Journal of Environmental Economics and Management* 7: 1-19.

Samuelson, Paul A. (1950). The problem of integrabilityin economic theory. *Economica* 17: 355-385.

Samuelson, William and Richard Zeckhauser (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty* 1: 7-59.

Shafer, Wayne and Hugo Sonnenschein (1975). Equilibrium in abstract economies without ordered preferences. *Journal of Mathematical Economics* 2: 345-348.

Thaler, Richard (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior and Organization* 1: 39-60.

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

Tversky, Amos and Daniel Kahneman (1991). Loss aversion in riskless choice: a reference-dependent model. *Quarterly Journal of Economics* 106: 1039-1061.

von Weizsäcker, Christian (1971). Notes on endogenous changes of tastes. *Journal of Economic Theory* 3: 345-372.

Notes

1. Tversky and Kahneman present their theory for preferences over bundles of two goods, saying that the generalisation to n goods is straightforward. We present an n-good version of the theory which we believe is faithful to their intentions. Where we have written that a move from r to s weakly (respectively: strictly) favours s, Tversky and Kahneman explicitly require only that $y \sim_r x$ implies $y \succeq_s x$ (respectively: $y \succ_s x$). But it is clear from their discussion that they also require that y

$\succ_r x$ implies $y \succ_s x$.

2. In personal communication, Kahneman has confirmed that, in developing their theory of reference-dependent preferences, he and Tversky began with the additive representation.

3. Tversky (1969) presents evidence of systematic intransitivities in preferences among lotteries, but in these cases, intransitivity is explained by the hypothesis that individuals ignore small differences in probabilities. This effect works in the opposite direction to diminishing sensitivity.

4. An alternative proof of Result 2 can be constructed by considering the additive representation, as defined by (1) in Section 2, with the additional assumption that each value function $v_i(.)$ is strictly concave. It can be shown that any preference structure that has such a representation satisfies C1–C7 and C7*. If (as Tversky and Kahneman usually assume) each $v_i(.)$ is kinked at the point at which $x_i - r_i = 0$, C7** is satisfied too.

5. Because reference-dependent indifference surfaces are smooth, C7** is *not* satisfied. Notice that, for all x, $\{z: z \; RR \; x\} = \{x\} \cup \{z: z \succ_L x\}$. Thus, moving in a sequence of infinitesimal steps, the individual is willing to make exchanges which take her away from her initial endowment, but keep her arbitrarily close to the long-run indifference surface on which she started.

6. We do not assume that the individual takes *full* advantage of such opportunities in each period. To do that would be to impose a restrictive synchronisation of trading decisions and shifts of reference points.