15 Incentive and information properties of preference questions: commentary and extensions

Richard T. Carson and Theodore Groves

INTRODUCTION

This chapter is both a commentary on and extension of Carson and Groves (2007) (hereafter CG) The substantial attention the paper has received has been enormously gratifying. Reception of CG has largely been positive with little if any substantive criticism directed toward it; and, there are many papers now being presented at conferences that are testing or relying on various aspects of it.

Our remarks are organized into a series of short sections. The first points out that the main purpose of CG was to extend the revealed preference paradigm to cover some types of survey responses. The second notes that CG provides the theoretical foundation that some critics of contingent valuation (CV) had argued was missing. The third takes the concepts of 'hypothetical' and 'hypothetical bias' head on and argues that these concepts are, for the most part, ill-defined or simply wrong and have done enormous damage to clear and careful thinking about the nature of the response to stated preference questions. The fourth examines the properties of cheap talk which is often proposed as a way to reduced hypothetical bias. The fifth provides some elaboration on CG and the issue of how to interpret information extracted from preferences questions. The sixth poses an answer to the often asked question: is a single binary discrete choice (SBC) question always the best elicitation format for a researcher to use? The seventh provides some elaboration on the payment card elicitation format, which in recent years has seen a resurgence. The eighth turns to an examination of some of the properties of the now widely used discrete choice experiment. The ninth considers the usefulness of economic experiments to help determine the performance of preference elicitation formats. The last section addresses the relationship between CG and the behaviouralist critique of neoclassical economics with a focus on the different-answers-to-the-same-underlying-question issue.

EXTENDING THE REVEALED PREFERENCE PARADIGM

The CG paper has not, of course, quelled objections by some economists to the use of preference information obtained from surveys to place monetary values on goods. Nor should it have. The paper's purpose was to suggest economists should think about surveys as a source of 'revealed' preference information. As long as the preference information collected in surveys is used by governments and private firms to help make decisions, then people *should* use the opportunity provided by their survey response to help influence those decisions. In this sense, responses to survey questions meeting the set of conditions CG term 'consequential' are no different than any other type of behaviour that economists use to infer information about preferences. One way to view CG is as just another evolutionary step along the path pioneered by Bowen (1943) who early on recognized that voting represented economic behaviour with respect to public goods and Becker (1978) who saw that the allocation of time between activities and even behaviour as intimate as marriage were reflective of underlying utility in the standard sense of neoclassical economics.

PROVIDING A THEORETICAL FRAMEWORK

Carson and Groves (2007) provide the underlying theoretical framework that Cummings and Harrison (1994, pp. 115–17) correctly pointed out was missing with respect to the use of the contingent valuation method (CVM):

> There exists no theory that relates to individual valuation behavior in markets or referendums under conditions in which the good being purchased or the issue on which people are to vote is hypothetical and implied economic commitments are hypothetical. Therefore, as a theoretical basis for applications of the CVM, one must presume that the received economic theory of individual behavior in markets where real economic commitments are made, or the majority rule principle derived in social choice theory, is relevant for the hypothetical context of the CVM. The consistency of people's valuation behavior in the CVM with that assumed in value theory or the majority rule principle is, of course, an empirical question. Unfortunately, there does not currently exist a body of empirical evidence that might establish this consistency in any compelling way. Thus there exists no basis for drawing unequivocal conclusions as to the theoretical substance of values derived with the CVM.

Carson and Groves (2007) provide this theoretical foundation by first dividing questions into two types, consequential and inconsequential. For a question to be consequential, survey respondents need to believe, at least probabilistically, that their responses to the survey may influence some decision they care about. For consequential survey questions, neoclassical economic theory is relevant in terms of the incentives respondents face in answering the question. Fortunately, most CV surveys fall into this category, as they usually ask about something the respondent cares about (even if it is only the possibility of increased taxes) and are clearly intended to be used as an input to some decision making process. We contend that it is implausible to believe that someone would go to the expense of conducting a survey if it were clearly a priori that the agency was going to ignore the information it supplies. Inconsequential questions are those for which there is either no chance of influencing a government or firm decision and/or when utility is not changed by the decision to be made. For inconsequential decisions, any response is as good as any other response in terms of its influence on the respondent's utility level. Inconsequential questions can easily be created in a laboratory situation but are harder to do so convincingly in an actual field survey. Thus, as emphasized by CG, it is inappropriate to lump all survey questions together and label them as 'hypothetical'. As CG note, the difficulty with the word hypothetical is that it is ill-defined, an issue to which we now turn.

HYPOTHETICAL SURVEY QUESTIONS AND HYPOTHETICAL BIAS

Critics of the use of stated preference surveys are quick with the word 'hypothetical' as a pejorative adjective in front of 'survey', 'question' or 'bias'. But what does the term hypothetical mean? Dictionary definitions include: (1) related to a hypothesis, (2) assumed or thought to exist, and (3) as a synonym for the logical term conditional, sometimes in the context of a conjecture in a legal situation. While this last definition is consistent with the use of the term 'contingent' as in a CV survey, the critical aspect to note is that *none* of these definitions explicitly embody the definition of having 'no influence' on a decision to be made. This, however, is the way the term is often used by economists when referring to surveys, and forms the basis of how most tests of hypothetical bias are operationalized in the experimental economics literature.[1]

In CG's framework these tests of hypothetical bias are tests of consequential versus inconsequential questions. Such tests, to be blunt, are *completely useless* in terms of determining the properties of consequential questions. As such, much of the discussion in the existing literature (for example, Murphy, et al., 2005) about hypothetical bias is misguided because much of the evidence from experimental tests is simply irrelevant.[2] What then is left? A very large meta-analysis (Carson et al., 1996) suggests that estimates based on contingent valuation are highly correlated with (and if anything, slightly smaller on average) than estimates based on household production functions and hedonic pricing. A substantial body of evidence re-examined every couple of years suggests that political polling on two candidate races and referendums taken close to elections by the standards of economic forecasting are excellent predictors on average of actual voting. A much smaller number of comparisons between referendum votes and directly comparable CV questions also find close correspondence between the two. There are repeated demonstrations that survey-based estimates of how much people would be willing to contribute voluntarily are substantially higher than actual contributions and indications that surveys tend to over-predict the purchases of newly introduced consumer goods.[3] This pattern of results is predicted by the CG framework. It is, however, not predicted if respondents always truthfully respond to preference questions in surveys or if there is always a substantial 'hypothetical bias' effect in these surveys.

The usual claim of widespread hypothetical bias in stated preference surveys comes from irrelevant experimental tests using an inconsequential treatment as the incorrect stand-in for a survey question or from field comparisons involving voluntary contributions or purchases of new private goods. That no predictions follow from neoclassical economic theory for inconsequential questions has already been noted.[4] The other two situations where claims of hypothetical bias tend to originate are both similar in nature in that positive survey responses may be reasonably expected to increase the likelihood that the good will be made available in the future and the agent would then have the option to get to contribute towards/purchase it at a later date. This provides an incentive for the respondents to overstate, which is what tends to be observed in practice, consistent with CG's prediction. There is also an incentive to free ride in terms of the actual contribution with respect to the public good while payment for the private good is necessary in the sense that it cannot be obtained without payment. The use of voluntary contributions to provide a public good provides the classic illustration of why the concept of hypothetical bias is ill-defined if not just simply wrong. Neoclassical theory suggests the survey should overestimate true willingness-to-pay (WTP) while actual contributions should underestimate true WTP. Why would anyone define the (hypothetical) bias of using the survey estimate as the difference between the estimated and the actual voluntary contribution?

CHEAP TALK

There is almost a complete disconnect between how the term cheap talk is used in the game theory literature and its use in the non-market valuation literature. From a theoretical perspective, cheap talk is an interesting communications concept first examined by Crawford and Sobel (1982) for games without a dominant strategy as a way of altering the nature of equilibrium strategies.[5] Ironically, this literature shows that talk is not 'cheap' when it can influence the actions of others. What was thought of initially as a costless way of signalling, in contrast to the original work on costly signalling by Spence (1974), turned out to be quite consequential in the right circumstances. Two parties with objectives that were not perfectly aligned or diametrically opposed might be able to use a 'cheap talk' signal to reach a mutually more advantageous outcome. While the cheap talk signal is costless to send, the economic value of the signal need not be zero and can be calculated for each party as the difference in economic value of the outcomes achieved with and without its use. As long as the signal has non-zero economic value, agents are not indifferent to its use. Cheap talk in the usual game theory context is not intended to alter the strategic incentives of a game nor does it supply any information about a particular agent's payoff options but rather yields information about the preferences of other agent(s) that is potentially useful for coordination on one or more equilibrium solutions.

The 'cheap talk' language in stated preference surveys tells respondents that some other respondents lie when they answer survey questions and, as such, the fraction of people who would actually vote in favour is smaller than the fraction that says yes in the survey (for example, Cummings and Taylor, 1999). This cheap talk script is an explicit attempt to introduce the notion of hypothetical bias. While this language was clearly inspired by the game theory theoretic literature on cheap talk, no one has ever laid out a formal economic model of how or why cheap talk should have an influence on survey responses.[6] Indeed, the use of the term cheap talk in the game theoretical literature, which is focused on the use of costless signals to helping the parties coordinate on more desirable outcomes, is sufficiently different from the use of term cheap talk in the non-market valuation literature so as to be a source of confusion.

Cheap talks' standard implementation in the non-market valuation literature has been with an SBC question for a public good, although not necessarily in a context that meets the CG conditions for incentive compatibility. We ignore that issue here and assume that the SBC question is incentive compatible, but that raises another question. An incentive-compatible SBC question has a dominant strategy response so it is unclear what role cheap talk is supposed to be playing. The stated purpose of cheap talk as explained by those implementing it in stated preference surveys is to reduce hypothetical bias. The problem is that anything that reduces WTP estimates tends to be seen as accomplishing this objective, but as we have pointed out earlier, the notion of hypothetical bias is ill-defined. Looking, for instance, at the main cheap talk script used in Cummings and Taylor (1999), the question that a rational respondent should ask is why is there a divergence between the survey response and the actual vote? There are many possible speculations, a lower quality good and the failure of some agents to follow through with actually paying. Unfortunately, these interpretations of the cheap talk script should have the effect of lowering the probability of a yes answer for some respondents. Because cheap talk should not have an impact unless it induces a change in the characteristics of the good or the payment obligation in situations where the agent has a dominant strategy, it should not be surprising that empirical tests of cheap talk produce erratic and inconsistent results (for example, Aadland and Caplan, 2006).

Parsing the language of different cheap talk scripts reveals 'hard' and 'soft' cheap talk versions.[7] The hard cheap talk version tells a respondent that some respondents lie when they say 'yes' in surveys. It is not clear how a respondent should interpret such a statement and it clashes with the usual social norm of truth telling that survey researchers try to advance in surveys. It is not hard to imagine interpretations that violate the standard notion of cheap talk and which should alter the answer that some respondents give. For instance, it is possible that some respondents see the statement as saying that other respondents had had buyer's remorse which might increase uncertainty over the characteristics of the good in a negative manner. Another interpretation is that some respondents might see this as a statement that other respondents were going to shirk their obligation to pay. This could, in turn, either decrease the likelihood of a particular respondent saying yes, either due to a reduced probability that the good would be provided if there were shirkers or through fairness considerations. It might also increase the likelihood of a yes answer if the respondent thought the cheap talk language indicated a possibility that people can shirk. No doubt one could argue over the plausibility or relevance of any of these interpretations or advance others, some of which may result in the appearance of a higher WTP. The point that CG make is that a researcher needs to consider seriously the informational content of survey statements and the influence they might have on respondents. The strong version of cheap talk simply has too many interpretations as to why it might have an effect and some of these interpretations lead to undesirable changes in the responses given to a survey question.

The soft version of cheap talk invokes the notion of a divergence between a casual survey response and what respondents would like to do if they carefully considered their budget situation. It differs from the strong version of cheap talk in that it does not invoke other respondents explicitly lying when answering survey questions. By focusing on individual failure to consider payment obligations as the source of the problem, it does not raise the issue of potential problems with the good. The soft version of cheap talk stresses the need for the person answering the survey to make sure that they could actually pay and recognize what the commitment they are making is. It has some similarities both with respect to content and intent to language used in some earlier contingent valuation studies that explicitly invoked the respondent's budget constraint and provided an opportunity to reconsider their response.

This soft version of cheap talk may reduce WTP estimates if there is a random component to respondent answers and it reduces that random component in an asymmetric manner.[8] Reducing the random component by inducing respondents to take more care in their answers is desirable from a policy perspective, although it is important to note that doing so can sometimes produce divergences with behaviour in actual markets where much less time and effort is often put into making decisions than in a survey context. The asymmetric nature of the cheap talk script by concentrating on the payment aspects rather than on the desirable quality of the good to be provided will tend to make WTP estimates more conservative.

When one moves away from an elicitation format in which a respondent has a dominant strategy to one that does not, then cheap talk can have an impact on a respondent's optimal answer. Finding a difference in estimates with and without using a cheap talk script should not be taken as an indication that hypothetical bias is present. A simple example extending an incentive-compatible SBC to a multinomial choice question with three alternatives illustrates the issue. Suppose a respondent is offered three choices A (the status quo option), B and C, where the agent's preferences are $C > A > B$. If the cheap talk script alters the respondent's perception of the fraction in favour of A in a downward direction, the respondent's optimal choice may now well be A, to avoid the worst option of B. This action clearly lowers the estimate of WTP for C, yielding the effect attributed to hypothetical bias. While the nature of the problem with cheap talk in this case is particularly easy to see, the ability of cheap talk to influence respondent beliefs about priors concerning particular goods and/or their attributes in more general settings where respondents do not have dominant strategies is conceptually straightforward to show. Hence, cheap talk can have an influence (potentially undesirable) on consumer choices in these situations both in surveys and actual markets. The issue of how information provided in a survey influences respondent beliefs concerning the preferences of other agents is one that naturally follows from the CG framework but it remains largely unexplored from an empirical perspective.

EXTRACTING PREFERENCE INFORMATION

Carson and Groves (2007) suggest that the interpretation of data from stated preference surveys is much more complicated than previously thought. Most researchers using such data had implicitly or explicitly assumed that people truthfully answered the questions they asked. Carson and Groves (2007) argue that in general, this assumption is likely to be false if the survey question is consequential and the respondent is acting like a rational economic agent. The conditions under which truthful preference revelation is always in the respondent's best interest are often hard and sometimes impossible to meet. However, one of the most important but often overlooked implications of CG is that even in the absence of incentive for truthful preference revelation, much useful information can be obtained from stated preference data. The key question that CG addressed was how to interpret such information and the nature of the deviations from truthful preference revelation that were likely to be observed in particular instances.

IS A SINGLE BINARY DISCRETE CHOICE ALWAYS THE BEST FORMAT?

Under specific conditions noted by CG, an SBC question for a pure public good is incentive compatible in the sense that truthful preference revelation is a respondent's dominant strategy. This, however, does not imply that stated preference surveys should always use this elicitation format. First, the incentive compatibility result requires auxiliary conditions to be met, and this is often difficult and sometimes impossible to do. This qualification is true of markets, voting and surveys. Second, when the SBC question being asked is incentive compatible, the preference information that it provides to the researcher is very limited. All a respondent's choice can do is increase or decrease the likelihood that the specified good is provided at the specified cost. This means that surveys using an SBC question require large samples and substantial pre-testing to help determine the range and placement of the cost amounts used. As such, a researcher may want to use an elicitation format that provides more preference information.

The main implication from CG on this point is that preference information from these alternative formats may be distorted. For instance, CG argue that a zero spike in the estimated WTP distribution is a natural consequence of the incentive structure induced by an open-ended matching question since a respondent should report a zero WTP if their true WTP is lower than the expected cost if the good is provided. There can, of course, in particular instances be alternative explanations for the observed phenomena that have a different interpretation. Many people could actually be indifferent to having the good supplied rather than simply having WTP values below expected cost.

There are two other aspects of an SBC question that are worth noting. The first is a corollary to the sparse nature of the information obtained from this elicitation format. Statistical precision of estimates either requires very large sample sizes and/or making strong assumptions about the nature of the underlying latent WTP distribution. Many of the claims about the SBC format overestimating WTP seem traceable to one of two problems: (1) making an inappropriate distributional assumption, and (2) inappropriate treatment of the right tail of the distribution. An example of the first problem is estimating a logit or probit model with the log of price as the stimulus variable. The second problem has cursed almost all welfare estimation irrespective of the source of the data obtained and has many variants. For travel cost models, functional form assumptions rather than actual data are relied upon to choke-off demand beyond some point. While stated preference questions can often push the prices for which reliable information can be obtained, both higher and lower than what is available in a market context, extremely high and extremely low prices should not be seen as plausible by respondents. As such, CG argue that rational respondents will answer the question with the cost information they think is relevant. The need for posing only realistic/credible questions cannot be emphasized enough. A more subtle variant of the right tail problem is the failure to recognize that data issues unrelated to the properties of the elicitation format may contaminate the data generating process. These include respondent confusion and the interviewer incorrectly recording the answer or the data entry incorrectly transferring the response. Note that these problems also occur with respect to data received from market transactions. In that context, they are often less obvious and less problematic as long as the highest price observed in the market is far from choking off demand or if the researcher is only interested in estimating marginal changes in WTP with respect to changes in an attribute rather than total WTP. The implications of consumer confusion with respect to both survey responses and market purchases is an issue that deserves more attention.

Second, a binary discrete choice question is most appropriate where there are two alternatives. A natural example is the status quo versus an alternative involving a public good (for example, status quo level of air quality in a city versus an alternative level). The critical feature is that only one level of air quality can be supplied so that any question that puts more than one alternative to the status quo into play unravels the SBC incentive properties. The issue is quite different when multiple additions to the status quo can be made available such as new fishing locations or new products. In this switch from public to quasi-public/private goods, the SBC question loses much of its attractiveness. The relationship between the nature of the good and the properties of the elicitation format developed by CG still appears to be under-appreciated.

A RESURGENCE OF THE PAYMENT CARD FORMAT

One of the more interesting developments in recent years has been the emergence of the payment card format first proposed in Mitchell and Carson's (1989) early work as the most popular matching elicitation format. The purest version of a matching elicitation format, the open-ended direct question, finds many respondents at a loss as to how to answer. At first this was thought to be related to asking about an unfamiliar public good but familiarity is not the main factor. In most western societies, making decisions in response to posted prices is the norm. In such a context, choice is the economic primitive which reveals preferences.

Interestingly, original criticism of the payment card was based on the anchoring behaviour with respect to the starting cost amount used, as seen in the bidding game format. Carson and Groves (2007) argue that this anchoring behaviour should be expected if the initially asked about cost is thought by the respondent to be correlated with the good's actual cost, which seems like a natural inference for respondents to make. Rather than encouraging a vague type of anchoring, the array of amounts on a payment card we conjecture may do two things. The first follows from the usual language of a payment card to pick any amount on the card.[9] This subtly converts the question into a choice question but one with a sufficiently large number of options that it ends up approximating a continuous matching response. The second is that the sequence of amounts on the payment card appears to increase uncertainty over the actual cost of the programme relative to the expectation that is formed in the open-ended direct question. Carson and Groves (2007) shows that increasing uncertainty with respect to cost in the matching format tends to increase the optimal stated WTP response towards its true value from below under most plausible belief structures. This may result in the payment card producing conservative WTP estimates, but not grossly conservative estimates. The theoretical drawback is that the payment card cannot be guaranteed to always provide incentives for revealing WTP amounts equal to or less than true WTP. In practice, there are usually only a small number of suspect very high WTP responses. Whether these responses are inconsistent with income levels and other covariates may be checked using regression procedures design to identify outliers.

CHOICE EXPERIMENTS

The increasingly popular discrete choice experiment (DCE) format received limited treatment in CG beyond a few key results. Carson and Groves (2007)'s starting point was to note that the SBC question, to which they devoted considerable attention, is the simplest case of a DCE. Moving from an SBC with two alternatives to a multinomial question with $k > 2$ alternatives generally causes a loss of incentive compatibility even if (1) the payment mechanism is coercive, (2) no other decision is potentially influenced by the response to the question, and (3) a take-it-or-leave it offer is made. The fundamental reason for this is that, if only one good is to be supplied, then a particular respondent's optimal choice should depend upon beliefs about the choices that are likely to be made by other respondents. As such, truthful preference revelation can no longer be a dominant strategy for all consumers and belief structures as it is in the case of SBC. When a survey's influence on the agency's decision comes through a plurality aggregation rule, for example, it is easy to show that, if all respondents have completely flat (that is, uninformative) prior assumptions about the choices likely to be made by other respondents, then truthful preference revelation is the optimal strategy.[10] The question for empirical researchers then is how likely is it for the flat prior assumption to hold? What is not known though without imposing a lot of structure on the problem is how a consumer should trade-off a weak but non-flat prior assumption against the strength of preference for a particular alternative. This is the situation that is likely to hold in most situations.

Carson and Groves (2007) also note that the truthful preference revelation problem in a multinomial choice question can go away in the special case where all but one of the $k-1$ of the goods rather than just one of good is provided.[11] It is easy to show in this situation that the multinomial choice question is effectively a SBC of the respondent's most preferred alternative paired against a single stochastically chosen less preferred alternative. This context is most likely to be applicable to quasi-public and private goods. Again, however, the situation facing empirical researchers is likely to be the intermediate case where there is uncertainty over the number of alternatives that might be provided.

Many DCE's utilize more than one choice set. This introduces a new issue. How does the agency aggregate responses across choice sets? Randomly picking one choice (under the assumption that respondents are expected utility maximizers) provides them with an incentive to treat each choice set as independent. While it is possible to provide assurances to participants in a laboratory experiment that this is what is being done, such a statement may not be credible in the context of a survey, as it suffers the same problem that CG point out occurs with any survey implementation of the Becker et al. (1964) mechanism. It may not be plausible to respondents that information collected in a survey would be discarded and not used. Most plausible aggregation rules result in a situation where the optimal response by some respondents to one choice set is contingent on the response they gave in another choice set.

A key insight of CG was that the nature of deviations from truthful preference revelation in non-incentive compatible DCEs should not manifest themselves as random behaviour. Most of the tests comparing consumer preferences estimated from DCEs to similar estimates using behaviour revealed in a market context are now based on whether the preference parameters from the two approaches are consistent, up to a constant scale (variance) factor.[12] Such tests, however, do not have much power against many forms of strategic behaviour as they are partially or completely confounded with changes in the scale factor. The objective of non-truthful revelation is to drive down price (for an existing good), to help induce provision (for a new good in the case where later purchase is an option) and/or to take account of the perceived preferences of other agents under some type of plurality decision rule. In none of these cases is random deviation from truthful preference revelation optimal.

The usual deviation from truthful preference revelation will be for respondents to sometimes indicate that their second most preferred option is the choice that they would make from the set available. A violation of the independence of irrelevant alternative (IIA) assumption results from this action.[13] A slightly different way to see the nature of the IIA violation is to note that the choice between any two alternatives now depends on the presence or absence of other alternatives. Independence of irrelevant alternative violations are typically seen in data from DCEs. It is straightforward to see how this type of IIA violation inflates the variance since the implicit variance has to increase to explain why the second favourite option was indicated as the choice out of the set of options. It is sometimes argued that strategic behaviour in a DCE is a difficult task for respondents to undertake, but all respondents have to do is to act as if they are more (or less) price sensitive than they actually are when the bundle of attributes they most prefer in a choice set is priced higher than they expected it be, given the other alternatives in the current or previous choice sets.[14] There are, of course, many other reasons posited for the ubiquitous IIA violations observed. The point we wish to make is that deviations from the standard conditional logit model now typically modelled as preference heterogeneity in a random parameters sense can also be generated by the sort of non-truthful preference revelation that one might expect to see in DCE.

The broader message from CG is that a researcher needs to step back and ask the question: what should a respondent answering a DCE be trying to accomplish? The most troubling answer is 'nothing', as this implies that the questions being asked are not consequential. By asking this question, CG provide an insight into a long-standing but little recognized puzzle. For private goods, choice questions tend to overestimate the propensity to buy a new product potentially being introduced into the market, while at the same time choice questions for existing products produce estimates that suggest survey respondents are more price sensitive than actual customers in stores. Neither result is surprising once one realizes that a respondent who potentially wants a new product to be available should act less price sensitive than they truly are to increase the likelihood of it being offered for sale in the market, while for an existing good, the same respondent should act more price sensitive in hopes of reducing the price the good is sold for.

Seeing different prices for the same or a closely related good can also influence a respondent's optimal answers to a sequential DCE. For instance, with a coercive payment mechanism for a pure public good, some respondents may rationally say 'no' if they have seen the same good or a closely related good earlier for a lower price.[15] There are other interpretations of what impact having the respondent seeing multiple prices for the same or closely related good can have. Take for instance the case of being offered the good at a higher price. The respondent may be more likely say 'no' even though their WTP exceeds the higher priced asked because they presume the good can be supplied at the first priced asked. This can lead to the appearance of starting point bias because answers to subsequent questions are 'anchored' on the first price seen, as it is deemed the most credible. A wide variety of different behaviours such as price averaging and completely ignoring very high or very low prices are plausible depending on how divergent information concerning prices is translated by the respondent into beliefs about what price will actually be paid if the good is supplied.

The general difficulty with a survey that presents the respondent with a sequence of choice sets is that the researcher would like the respondent to treat each choice set as independent of the other, but there may be no reason for a rational respondent to do so. Failure to treat the choice sets as independent could be manifest in any number of ways but one way (Day et al., 2009) appears to be for the respondent to accept the attribute levels other than cost and then to adjust the perceptions of the actual cost to be paid (including uncertainty about cost in the case of a coercive payment mechanism). This can be seen in respondents who either become much more or much less price sensitive than they would be in an incentive compatible SBC or the actual marketplace. The particular effect that should be expected depends on the nature of the payment obligation and how respondents believe the agency will use the information with respect to price (or other attributes).[16]

THE USEFULNESS OF ECONOMIC EXPERIMENTS

Economic experiments, both laboratory and field, have the potential to shed considerable light on the incentive and informational properties of survey elicitation formats. Their track record to date, however, has been quite mixed. Much of the problem stems from an obsession to show whether hypothetical bias exists, which is an understandable research endeavour given the scepticism many economists have concerning the use of surveys.

The most common experiment has been a blunt instrument testing the hypothesis that respondents always tell the truth, irrespective of the incentives they face for preference revelation, rather than a test of any theoretical prediction from economic theory. A well-known and widely cited example is Cummings et al. (1995). This paper compares the percentage who say ‘yes’ that they would pay a specified amount for various private goods when payment is required to the percentage who say ‘yes’ in a treatment where it is made clear that the question being asked is purely hypothetical in the sense that they will neither pay for nor receive the good. They find that more subjects in the ‘real’ treatment say ‘yes’ than in the ‘hypothetical’ treatment. This result has led some researchers to believe that contingent valuation overestimates and is frequently invoked by critics as a reason why contingent valuation methods should not be used. But consider this experiment through the lens of CG. If respondents considered the second treatment to be inconsequential, then economic theory makes no prediction with respect to comparing the two treatments.[17] It may be useful to ask the question: what results would the researcher have expected to see if statements made in the second treatment that the response was inconsequential were ignored? In that case, anyone who thought that they might want to accept a future offer to obtain the good should say 'yes' because saying 'yes' might increase the likelihood that such an offer would be made and the respondent could decided at that point whether to accept it.[18] Thus, if the second treatment was taken by respondents to be inconsequential, then it is not clear why the comparison is of any interest to economists. If the treatment was taken as consequential, then the theoretically predicted result was observed for a private good was observed.

Carson and Groves (2007) has inspired a substantial amount of experimental work and it is beyond the scope of our effort here to comprehensively review it, although we believe that in a few years from now this would be a worthwhile endeavour. Carson et al. (2004) provide experimental results supporting a key implication of the CG framework: the probability of a vote on a public good being binding does not influence the fraction responding 'yes' as long as it is positive. There are two other findings from this study. First, a purely hypothetical case (probability of the vote being binding is zero) does behave differently from treatments where the probability of the vote being binding is positive. The empirical estimates are overestimates which is consistent with past experimental work, suggesting that results from the purely hypothetical case should not be relied upon as an indication of how consequential questions work.[19] Second, creating an explicit linkage between two decisions influences the response to a question asked about only one of those decisions. While the nature of the linkage was made obvious to subjects, the response to it was clearly inconsistent with the belief that respondents always truthfully reveal their preferences and provides a note of caution with respect to making the assumption that respondents are not capable of linking multiple issues.

Some of the most interesting papers we have seen have taken a step away from the generic (and, as CG argue, ill-defined) question of whether contingent valuation overestimates, to look at the nuts and bolts of how specific elicitation formats work under much more controlled circumstances using induced rather than home-grown preferences.[20] Doing so allows clearer identification of deviations from theoretical predictions and the ability to better sort out what type of belief structures subjects are using in particular contexts. One of the key findings that emerges in this work is that subjects make optimization errors, an issue that CG do not consider. Taylor et al. (2001) show that optimization errors using a referendum format requiring payment are relatively common, even though the aggregate results were consistent with what one would have expected under truthful preference revelation. Using induced values, Palomé (2003) shows that about half of the subjects responded truthfully and that those who did not respond truthfully were much more likely to under-declare (41 per cent) versus over-declare (12 per cent). Vossler and McKee (2006) look at several different elicitation mechanisms and show that they differ in terms of the fraction of subjects making optimization errors. An intriguing result in this paper is that asking subjects about how certain they are about their answers induced the appearance of uncertainty even where there was none. Collins and Vossler (2009) used induced preferences and looked at the difference in choices made by respondents when faced with three options. They create a context in which a subject should view all three of the options as having an equal probability of being chosen by other agents. In situations involving equal prior assumptions with respect to the choice of other agents, CG argue that the multinomial choice question with three options is incentive compatible. Collins and Vossler's overall results are consistent with truthful preference revelation, albeit with optimization errors. Interestingly, treatments that move away from easy to understand plurality rules for determining outcomes toward more complicated schemes generally appear to induce a higher fraction of optimization errors. The lesson from all these studies is that errors of optimization are likely to be more common than often thought and can vary with the nature of the task. Theoretical work on the nature and implications of such optimization errors is clearly needed, as most random utility models ascribe the error component to specification error and factors observable to the agent but not the analyst.

The induced value framework can also be used to examine other key issues related to the properties of elicitation. For instance, Carson et al. (2009) show that one can achieve incentive compatibility in a double-bounded discrete choice question if the link between the two questions can be broken. This is easy to do in a laboratory experiment but hard to do in an actual field survey. By tracing out the steps involved, they are able to isolate the source of the problem which is the ability to guarantee the independence of the two questions. This, in turn, may provide some insight into situations where the independence of the two questions is more likely to approximately hold.

BEHAVIOURALIST CRITIQUE

The CG paper presents a set of neoclassical predictions that stand in stark contrast with several of the key predictions associated with the behavioral critique of neoclassical economics. Chief among these is the assertion that the core problem with the assumption of neoclassical economic behaviour (that is, Tversky et al., 1990) is that it should produce estimates that are 'procedurally invariant', where the classic example are the divergences in the implicit preferences suggested by answers to choice and matching questions. Carson and Groves (2007) suggests that observing such procedural invariance under neoclassical economic behaviour is highly unlikely, as most approaches to obtaining preference information differ either in terms of their incentive structure with respect to truthful preference revelation and/or with respect to the nature of the information that a procedure conveyed. Carson and Groves (2007) predict procedural invariance will be violated and often provides guidance on the direction of the divergence that should be observed. A broad array of empirical evidence supports these predictions.[21]

Carson and Groves do not claim that neoclassical economic theory is not vulnerable to the behavioural critique. We do believe, though, that to the extent the behavioral critique is valid, it is unlikely to have different implications for consequential surveys than other types of economic behaviour. In this sense, it is inconsistent to act as if there are behavioral economics related problems with surveys used for valuation purposes that do not permeate data used to infer preference information obtained from other sources. The main distinction would appear to be that in a survey context it is easier to run experiments to examine the role of different types of effects. Given the ability of surveys to frame questions for respondents that avoid some decision-making problems and facilitate transparency, one might even argue that preference information elicited from well constructed surveys should play a larger role in helping to formulate policies that increase social welfare.

One of the most interesting directions for future research we believe is how the CG neoclassical framework and various predictions from behavioral economics interact. At the heart of CG's reading of the empirical evidence is that neoclassical marginal conditions appear to hold while much of the behavioural critique concerns stepping back to a much more primitive level regarding behaviour. Bernheim and Rangel (2009) provide an examination of what welfare economics might look like if it is based on consumer choice which is influenced by factors identified by the behavioural critique.

NOTES

1. This definition is surprising given Samuelson's (1954) early recognition that if the government relied on questionnaires of the public for preference information, it would be in the interest of respondents to exploit that opportunity for their own selfish gain.

2. Murphy et al. (2005) presciently state that their 'results are quite sensitive to model specification, which will remain a problem until a comprehensive theory of hypothetical bias is developed.'
3. It is worth noting here that the cleanest recent comparison in the literature to the ideal situation put forth by CG is Johnson (2006) and in Chapter 9 of this volume. This study looks at a case were a survey with a binary discrete choice question on a water supply issue in Rhode Island was first administered as an input to the policy-making process and then a subsequent binding referendum vote was held at one of the price points. At the $250 price used in the binding referendum, the percentage in favour was 46 per cent while in the survey the percentage in favour was 48 per cent, with the difference not statistically significant (p = .69).
4. Inconsequential questions in many contexts are odd in that they invite speculation as to their potentially hidden purpose. In many experimental contexts participants may believe that if they indicate responses consistent with higher (or in some contexts lower) WTP amounts that they will be more likely to be given the opportunity to participate in subsequent rounds where real money can be earned. This possibility makes the interpretation of the result of inconsequential treatments in economics experiments even more difficult to interpret.
5. See Farrell and Gibbons (1989) for further theoretical development and Farrell and Rabin (1996) for a very readable review.
6. Cummings and Taylor at the end of their paper suggest potential psychological explanations related to priming.
7. The environmental valuation literature (for example, Aadland and Caplan, 2006) sometimes refers to long and short versions of cheap talk that have some overlap with but do not directly correspond to the hard and soft labels we use here.
8. Interestingly, Bulte et al. (2005) comparing treatments using a cheap talk script to those emphasizing the consequential nature of the survey find no difference in the aggregate estimates. In contrast, treatments that appear very hypothetical tend to result in higher WTP estimates.
9. Most work on payment cards has focused on whether restricting the range of the amounts shown on a payment card can influence estimated WTP (Covey et al., 2007; Dubourg et al., 1997; Rowe et al., 1996) to which there are mixed results. Under CG, the range of amounts displayed on a payment card can have an influence on both expected cost and the level of uncertainty surrounding that amount. Earlier researchers (for example, Mitchell and Carson, 1989) looked at placing the cost of other goods (public and private) on the payment card. These, too, can have an influence on formation of cost expectations. Dubourg et al. (1997) demonstrate a pure psychological anchoring effect by giving half the sample a payment card starting with low numbers (and increasing) and the other half the sample a payment card starting with high numbers (and decreasing). This may be an indication that one of the approaches has violated standard conversation conventions (Grice, 1975).
10. Note that this is a sufficient condition rather than a necessary condition, so an empirical researcher is not able to conclude that a choice question is not incentive compatible if respondents do not hold a flat prior assumption.
11. This holds under the maintained assumption that the respondent gets utility from at most one of the goods in the choice set. Otherwise, one has to take account of the relationship between goods in the set and their joint consumption.
12. Choice models inherently produce estimates of (β/σ) rather than β, where β is the preference parameter of interest and σ is the scale factor. Some earlier tests comparing stated and revealed preference data made the mistake of comparing parameter estimates implicitly assuming that the scale factor in the two types of data was the same. Swait and Louviere (1993) and Haab et al. (1999) show that one needs to take account of the difference in scale factors in between survey and market/experimental data in comparing estimates since discrete choice models inherently generate parameter estimates confounded with a scale factor. In a particular context, survey data may be associated with a smaller

or larger random component than data from other sources. Further, the particular elicitation format and specific features of it, such as the number of attributes and/or attribute levels, may influence the magnitude of the random component. In spite of calls (for example, Louviere et al., 2002) for more research on factors influencing the magnitude and nature of the random component, there is still too little work on this topic relative to work concentrating on factors that might influence location shifts in parameters.

13. To see this, consider the case where there are three options (A, B and C), rank ordered for convenience in terms of a respondent's preferences. A is thought to have little chance of being implemented given beliefs about other agent's preferences but is close to B in attribute space with B being strongly preferred to C. At the heart of IIA violations is a dependence on the choice between two alternatives in the presence of one or more alternatives in the choice set.
14. If respondents adopt this simple strategy, then as CG argue, it is possible to correctly recover marginal WTP estimates with respect to changes in attribute levels. This is because the biased price/scale effect cancels out in the standard approach to obtaining marginal WTP estimates. Most comparisons in the literature (for example, Carlsson and Martinsson, 2001) look at marginal rather than total WTP. One of the few papers to look at both is Lusk and Schroeder (2004). They found in the context of a private good that marginal WTP estimates are similar between stated preference and experimental treatments while the total WTP estimates differed.
15. For an example documenting such behaviour, see Day et al. (2009). For private goods, this type of effect could go either way depending upon whether the respondent was more interested in influencing the price of the good or the probability that it was offered for sale.
16. An under-appreciated issue in surveys briefly raised by CG and explored indepth (Corrigan et al., 2008; Zhao and Kling, 2001, 2004) is the issue of *when* a consumer is making an irrevocable commitment to pay for a good if supplied. We suspect that some of the observed differences in WTP (and WTA) are really the result of subtle divergences between treatments on this dimension of 'commitment dynamic' that is inferred by respondents/experiment participants but not recognized by the researchers involved.
17. The general difficulty we see with the 'pure hypothetical' implementation in many experiments is its lack of plausibility. With money being spent to gather data from them, subjects should speculate as to the use that data will be put and respond accordingly. As long as some subjects believe that saying 'yes' or giving high WTP amounts will make more likely that they will advance, in the sense of being made future offers, then there may be an intrinsic tendency of purely hypothetical treatments to overestimate.
18. The issue with the experiment from this perspective is the assumption that public and private goods had the same theoretical properties. This assumption appears to have stemmed from the belief that if an elicitation procedure is not 'well-behaved' with private goods then it will not be well-behaved with public goods. This belief is pervasive in the literature even though it has no basis in neoclassical economic theory.
19. While this overestimate result may be an empirical regularity, without any theoretical basis it is unclear why such an empirical regularity exists.
20. This comment should not be taken as suggesting that home-grown values (that is, preferences that are not induced by the experimenter) should never be used, but rather, more caution should be taken when they are. The researcher should first ask whether the question can be best addressed using induced values. Herriges et al. (2007) and Vossler and Evans (2009) provide interesting experiments that explore the implications of CG's consequentiality using home-grown values. Both papers find similar results in quite different contexts. The concluding sentence of the Herriges et al. (2007) paper's abstract provides a nice summary: 'We find evidence consistent with the knife-edge theoretical results, namely that the willingness to pay distributions are equal among those believing the survey to be at least minimally consequential, and divergent for those believing that the survey is irrelevant for policy purposes.' We believe that more work on how to best induce consequentially in preference surveys is clearly needed.

21. We have given short treatment to the information aspects of the CG framework. Most experiments and surveys comparing different treatments have implicitly assumed that all information provided is taken at face value and clearly understood. A good example is the Powe and Bateman (2004) study which looks at the well-known external scope test (Carson and Mitchell, 1993, 1995) using flood protection schemes involving different parts of the Broadlands area in the UK. Their results suggest respondents are not sensitive to enacting the scheme for specific areas and for the entire Broadlands, a troubling finding. However, 41 per cent of respondents do not consider a scheme involving the whole area realistic. This fraction is much higher than the fraction finding the scheme unrealistic for specific areas. Not seeing a scheme as realistic is closely tied to not being willing to pay anything for it, a behaviour one might expect of rational agents. After controlling for whether the respondent sees particular schemes as realistic, the theoretically expected result that WTP for the more inclusive area is larger is now obtained. Their simple take home message is that performing like-for-like scope tests is harder than it seems.

REFERENCES

Aadland, D. and A.J. Caplan (2006), 'Cheap talk revisited: new evidence from CVM', *Journal of Economic Behavior and Organization*, **60**, 562–78.

Becker, G.M., M.H. DeGroot and J. Marschak (1964), 'Measuring utility by a single response sequential method', *Behavioral Science*, **9**, 226–32.

Becker, G.S. (1978), *The Economic Approach to Human Behavior*, Chicago, IL: University of Chicago Press.

Bernheim, B.D. and A. Rangel (2009), 'Beyond revealed preference: choice theoretic foundations for behavioral welfare economics', *Quarterly Journal of Economics*, **124**, 51–104.

Bowen, H.R. (1943), 'The Interpretation of Voting in the Allocation of Economic Resources', *Quarterly Journal of Economics*, **58**, 27–48.

Bulte, E., S. Gerking, J.A. List and A. de Zeeuw (2005), 'The effect of varying the causes of environmental problems on stated WTP: evidence from a field study', *Journal of Environmental Economics and Management*, **49**, 330–42.

Carlsson, F. and P. Martinsson (2001), 'Do hypothetical and actual marginal willingness to pay differ in choice experiments?', *Journal of Environmental Economics and Management*, **27**, 179–92.

Carson, K.S., S.M. Chilton and W.G. Hutchinson (2009), 'Necessary conditions for incentive compatibility in double referenda', *Journal of Environmental Economics and Management*, **57**, 219–25.

Carson, R.T. and T. Groves (2007), 'Incentive and information properties of preference questions', *Environmental and Resource Economics*, **37**, 181–210.

Carson, R.T., T. Groves and J. List (2004), 'Probabilistic influence and supplemental benefits: a field test of the two key assumptions underlying stated preferences', paper present at NBER Public Economics Workshop, Palo Alto, March.

Carson, R.T., N.E. Flores, K.M. Martin and J.L. Wright (1996), 'Contingent valuation and revealed preference methodologies: comparing the estimates for quasi-public goods', *Land Economics*, **72**, 80–99.

Carson, R.T. and R.C. Mitchell (1993), 'The issue of scope in contingent valuation studies', *American Journal of Agricultural Economics*, **75**, 1263–7.

Carson, R.T. and R.C. Mitchell (1995), 'Sequencing and nesting in contingent valuation studies', *Journal of Environmental Economics and Management*, **28**, 155–73.

Collins, J.P. and C.A. Vossler (2009), 'Incentive compatibility tests of choice experiment value elicitation methods', *Journal of Environmental Economics and Management*, **58**, 226–35.

Corrigan, J.R., C.L. Kling and J. Zhao (2008), 'Willingness to pay and the cost of commitment: an empirical specification test', *Environmental and Resource Economics*, **40**, 285–98.

Covey, J., G. Loomes, and I.J. Bateman (2007), 'Valuing risk reductions: testing for range biases in payment card and random card sorting procedures', *Journal of Environmental Planning and Management*, **50**, 467–82.

Crawford, V.P. and J. Sobel (1982), 'Strategic information transmission', *Econometrica*, **50**, 1431–51.

Cummings, R.G. and G.W Harrison (1994), 'Contingent valuation', in R.A. Eblen and W.R. Eblen (eds), *Encyclopedia of the Environment*, Boston, MA: Houghton Mifflin, pp. 115–17.

Cummings, R.G., G.W. Harrison and E.E. Rutström (1995), 'Homegrown values and hypothetical surveys: is the dichotomous choice approach incentive compatible?', *American Economic Review*, **85**, 260–66.

Cummings, R.G. and L.O. Taylor (1999), 'Unbiased value estimates for environmental goods: a cheap talk design for the contingent valuation method', *American Economic Review*, **89**, 649–65.

Day, B., I.J. Bateman, R.T. Carson, D. Dupont, J.J. Louviere, S. Morimoto, R. Scarpa and P. Wang (2009), 'Task independence in stated preference studies: a test of order effect explanations', CSERGE working paper EDM 09-14, Centre for Social and Economic Research on the Global Environment, University of East Anglia.

Dubourg, W. B., M.W. Jones-Lee, G. Loomes (1997), 'Imprecise preferences and survey design in contingent valuation', *Economica*, **64**, 681–702.

Farrell, J. and R. Gibbons (1989), 'Cheap talk can matter in bargaining', *Journal of Economic Theory*, **48**, 221–3.

Farrell, J. and M. Rabin (1996), 'Cheap talk', *Journal of Economic Perspectives*, **10**, 103–18.

Grice, H. (1975), 'Logic and conversation', in P. Cole and T. Morgan (eds), *Syntax and Semantics: Vol. 3, Speech Acts*, New York: Seminar Press.

Haab, T.C., J.C. Huang and J.C. Whitehead (1999), 'Are hypothetical referenda incentive compatible? A comment', *Journal of Political Economy*, **107**, 186–96.

Herriges, J., C.L. Kling, C.C. Liu and J. Tobias (2007), 'What are the consequences of consequentially', paper presented at Allied Social Sciences Meeting, Chicago, January.

Johnson, R.J. (2006), 'Is hypothetical bias universal: validating contingent valuation responses by a binding referendum', *Journal of Environmental Economics and Management*, **52**, 469–81.

Louviere, J.J., D. Street, A. Ainslie, T.A. Cameron, R.T. Carson, J.R. DeShazo, D. Hensher, R. Kohn and T. Marley (2002), 'Dissecting the random component', *Marketing Letters*, **13**, 177–93.

Lusk, J.L. and T.C. Schroeder (2004), 'Are choice experiments incentive compatible? A test with quality differentiated beefsteaks', *American Journal of Agricultural Economics*, **86**, 467–82.

Mitchell, R.C. and R.T. Carson (1989), *Using Surveys to Value Public Goods: The Contingent Valuation Method*, Washington, DC: Resources for the Future.

Murphy, J.J., P.G. Allen, T.H. Stevens and D. Weatherhead (2005), 'A meta analysis of hypothetical bias in stated preference surveys', *Environmental and Resource Economics*, **30**, 313–25.

Palomé, P. (2003), 'Experimental evidence on deliberate missrepresentation in referendum contingent valuation', *Journal of Economic Behavior and Organization*, **52**, 387–401.

Powe, N.A. and I.J. Bateman (2004), 'Investigating insensitivity to scope: a split sample test of perceived scheme realism', *Land Economics*, **80**, 258–71.

Rowe, R., W.D. Schuzle and W.S. Breffle (1996), 'A test for payment card bias', *Journal of Environmental Economics and Management*, **31**, 178–85.

Samuelson, P.A. (1954), 'The pure theory of public expenditure', *Review of Economics and Statistics*, **36**, 387–89.

Spence, M. (1974), *Market Signaling: Informational Transfer in Hiring and Related Screening Processes*, Cambridge, MA: Harvard University Press.

Swait, J and J.J. Louviere (1993), 'The role of the scale parameter in estimation and comparison of multinomial logit models', *Journal of Marketing Research*, **30**, 305–14.

Taylor, L.O., M. McKee, S.K. Laury and R.G. Cummings (2001), 'Induced value tests of the referendum voting mechanism', *Economic Letters*, **71**, 61–5.

Tversky, A., P. Slovic and D. Kahneman (1990), 'The causes of preference reversals', *American Economic Review*, **80**, 204–17.

Vossler, C.A. and M.F. Evans (2009), 'Bridging the gap between the field and the lab: environmental goods, policy maker input, consequentiality', *Journal of Environmental Economics and Management*, **58**, 338–45.

Vossler, C.A. and M. McKee (2006), 'Induced value tests of contingent valuation elicitation mechanisms', *Environmental and Resource Economics*, **35**, 137–68.

Zhao, J. and C.L. Kling (2001), 'A new explanation for the WTP/WTA disparity', *Economic Letters*, **73**, 293–300.

Zhao, J. and C.L. Kling (2004), 'Willingness to pay, compensating variation and the cost of commitment', *Economic Inquiry*, **42**, 503–17.