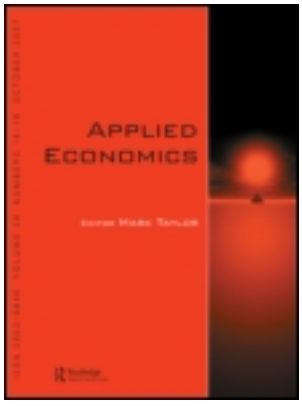

Applied Economics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/raec20

Estimating willingness to pay: do health and environmental researchers have different methodological standards?

Semra Özdemir [a] & F. Reed Johnson [b]

[a] Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27517, USA

[b] RTI International, 3040 Cornwallis Road, P.O. Box 12194, Research Triangle Park, NC 27709-2194, USA

Available online: 05 Apr 2012

To cite this article: Semra Özdemir & F. Reed Johnson (2013): Estimating willingness to pay: do health and environmental researchers have different methodological standards?, Applied Economics, 45:16, 2215-2229

To link to this article: http://dx.doi.org/10.1080/00036846.2012.659345

Applied Economics, 2013, **45**, 2215–2229

Routledge
Taylor & Francis Group

Estimating willingness to pay: do health and environmental researchers have different methodological standards?

Semra Özdemir[a,*] and F. Reed Johnson[b]

[a]*Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27517, USA*
[b]*RTI International, 3040 Cornwallis Road, P.O. Box 12194, Research Triangle Park, NC 27709-2194, USA*

Health and environmental economists have been employing Stated-Preference (SP) methods such as conjoint analysis or contingent valuation to estimate the monetary value of public health interventions and environmental goods and services. However, the quality of data and the validity of results are sensitive to a number of decisions researchers make. The aim of this study is to compare the degree of the current consensus among active researchers in the rapidly evolving area of SP methods in health and environmental valuation. We surveyed researchers who have published manuscripts on SP methods in the last 10 years. Researchers were presented with hypothetical SP studies with different attributes. They were first asked which study they would recommend to use to inform policy decisions, and then asked which study has better-quality. Our results show that good-practice SP methods vary among study features and among researchers with different amounts and kinds of research experience. Although health researchers had specific preferences on which study features were better, their quality judgements were not very consistent with their judgements about the acceptability of studies for policy analysis. On the other hand, environmental researchers had similar preferences over the study attributes for the two types of questions.

Keywords: stated preferences; researcher preferences; willingness to pay; health; environment

JEL Classification: I10; Q51

I. Introduction

New and better methods constantly are being developed in applied research. Peer review assumes a common and constantly evolving consensus among active researchers on acceptable research methods. A reasonable degree of consensus is the basis for review of scientific manuscripts for publication. Eventually, state-of-the-art procedures become state-of-the-practice procedures that are required for peer-reviewed publications. Defining state of the practice especially is important for Stated-Preference (SP) studies, where results are highly sensitive to researcher judgements at various stages of survey development and data analysis (Johnson and Desvousges, 1997; McIntosh and Ryan, 2002; Lloyd, 2003; Ryan and Amaya-Amaya, 2004; Lancsar and Louviere, 2006).

Market researchers, environmental economists, and transportation economists began employing such SP methods as Contingent Valuation Methods (CVM) and Attribute-Based Methods (ABM) (also known as discrete-choice experiments, conjoint analysis or stated-choice surveys) in the 1970s (Green

*Corresponding author. E-mail: semra@live.unc.edu

and Rao, 1971; Randall *et al*., 1974; Green and Wind, 1975; Mitchell and Carson, 1989). Researchers developed SP methods because of the need to obtain monetary values for goods and services for which there are no market prices, or for which market prices are a poor measure of societal values. These methods have been widely used in benefit-cost analysis for resource-allocation decisions (Boyle, 2003; Hanley *et al*., 2003) and have been tested for internal and external validity (Bryan *et al*., 2000; Shiell *et al*., 2000; Carlsson and Matinsson, 2001; Ryan and Bate, 2001; Miguel *et al*., 2002, 2005; Ryan and Miguel, 2003; Schwappach and Strasmann, 2005).

The SP approach is designed to measure the value of goods or services using subject evaluations of one or more hypothetical scenarios or alternatives. Designing such a study requires numerous judgements involving such considerations as what kind and how much information needs to be provided to respondents to prepare them for the evaluation task, what features of the outcomes to include, how many alternatives to show, how many groups of alternatives to show, how to construct alternatives to satisfy particular statistical criteria, and how to estimate preference parameters to obtain valid, unbiased results.

Over time, significant advances in nonmarket valuation have been proposed and adopted by researchers (Carson *et al*., 1998; OMB, 2003). Nevertheless, the role that various survey-design features play in eliciting valid preference measures is still a matter of debate and continues to be a topic of active research. Because survey methods in this area continue to evolve, a simple review of published studies may not indicate active researchers' current perception of good research practice.

While these methods have wide acceptance among both academic researchers and policy makers to value environmental goods and services (Johnson and Desvousges, 1997; Bennett and Adamowicz, 2001; Olsen and Smith, 2001; Bateman *et al*., 2002; Holmes and Adamowicz, 2003), valuing monetary benefits to assist in allocating health resources is less widely accepted. Nevertheless, health economists increasingly are adopting valuation methods developed previously in environmental economics and market research (Diener *et al*., 1998; Johnson *et al*., 1998; Ryan *et al*., 2001; McIntosh *et al*., 2010).

The objective of this study is to compare the degree of current consensus in applications of SP research in health and environment valuation. We administered an SP survey to elicit researchers' judgements for the methods themselves. This SP survey asked active researchers in the field to evaluate alternative hypothetical studies defined by attributes describing survey design and data analysis.

There are several reasons why we might expect differences in how environmental and health researchers conduct SP research. First, some differences might arise because of the nature of the field. Environmental applications generally value public goods, while health applications generally value private goods.[1] Second, legislative mandates and regulatory guidelines of the Environmental Protection Agency (EPA) and the Office of Management and Budget in the US and Department of Environment, Food and Rural Affairs in the UK reflect a general consensus among environmental economists regarding acceptable research methods. However, quantifying health benefits in monetary terms has limited acceptance in general, and both National Institute of Health (NIH) in the US and National Institute for Clinical Excellence (NICE) in the UK recommend using cost-effectiveness and cost-utility analysis in health applications (Drummond *et al*., 2005). Third, there may be differences among the researchers in the two fields because of their background and training. Environmental researchers generally have formal nonmarket-valuation training in economics or agricultural economics programs, whereas backgrounds of health researchers are more diverse. Although health researchers have borrowed SP methods from environmental researchers, the overall effect of the differences in the objects of valuation, institutional context, and/or academic training and disciplinary culture of the researchers may result in different perceptions of good SP practice in environmental and health research.

This study describes the characteristics of active environmental and health SP researchers, quantifies the range of consensus on good-practice methods among both groups of researchers, and identifies important areas of disagreement between the two groups. Our results also help identify where the SP research stands in health applications and how the current practice differs from the environmental research where these methods have been developed.

II. Methods

Survey development

An initial list of method attributes and levels and draft of the survey were developed based on a review of the literature and consultations with senior SP researchers. We then conducted a pilot study of the survey with active environmental-valuation researchers. An initial draft of the environmental version of the web-enabled survey instrument was administered to the W1133[2] members and the results of this pilot study were presented at the W1133 Annual Meeting in February 2005 in Salt Lake City. Based on data and comments we received from this group of environmental and resource economists, we revised the list of attributes and levels used to describe the hypothetical study profiles and changed the elicitation format.

After the first round of revisions, the draft of the survey instrument was pretested with 16 active researchers in environmental or health fields. Of these one-on-one interviews, 13 were conducted in-person and three were conducted on the telephone. The participants represented a wide range of backgrounds from junior to senior researchers, working in academic, industry and nongovernmental organization settings. Most participants had experience with different types of

[1] We acknowledge that this is a generalization. Environmental researchers may study private services such as recreational activities or eco-tourism, while health researchers may study public goods such as health effects of air pollution.
[2] W1133 is a regional research group working on benefits and costs of natural resources policies affecting public and private lands in the US. The environmental, resource and agricultural economists from the US universities are members of this group.

Table 1. Study attributes and levels

Feature		Options	
Question format		CVM: • Open-ended • Single or double bounded dichotomous choice • Dichotomous choice with uncertainty or other follow-up	ABM: • Ranking or rating of scenario list • Rated or graded pairs • Choice format
Experimental design		CVM: • Bid structure based on literature review • Bid structure based on a pilot study	ABM: • D-efficient design • Catalog-based design
Survey development		• Included focus groups, pre-test interviews and reviewed by technical experts • Included pre-test interviews • Included no focus groups or pre-test interviews	
Administration mode (using comparable representative samples)		• Mail • Telephone • In-person interview • Web-based	
Analysis	Scope test*	• Pass test • No test • Fail test	
	Income significant determinant of WTP	• Yes** • No	
	Scenario rejection, outliers, inconsistency, 'don't know' responses	• Adjustment • No adjustment	
	Statistics	• Basic modelling (such as OLS, logit or probit) • Advanced modelling (such as nonparametric modelling or mixed logit)	

Notes: OLS – Ordinary Least Square.
*Test of whether subjects' values vary positively and commensurately with an increasing quantity of the good offered.
**Yes: The study confirmed that income is a statistically significant determinant of WTP. No: The study did not explore whether income is a statistically significant determinant of WTP.

SP methods, but we also included researchers who had direct experience with only one of the question formats. The participants were encouraged to 'think aloud', expressing their reactions as they completed the survey questions. While information from these pretests was largely supportive of the initial draft survey, minor adjustments were made to the content and ordering of the text in the final survey to improve comprehension.

The attributes that were used to describe the scenarios in the final survey instrument include question format, experimental design, survey development, administration mode and an analysis section that included the use of a scope test[3]; whether income was a significant determinant of Willingness To Pay (WTP)[4]; handling of scenario rejection, outliers, inconsistency, and 'don't know' responses; and statistical methods. The question formats for CVM studies included open-ended, single or double-bounded dichotomous choice, and dichotomous choice with uncertainty or other follow up. The question formats for an ABM study were choice, ranking and rating/graded pairs. The bid structure in a CVM study was designed either based on a pilot study or literature review. The experimental design in an ABM study was either a catalog design or a D-efficient design. The rest of the attributes defined levels that can be used in either method. Table 1 lists the eight attributes and the levels selected for the survey questions. We provided explanations for all attributes and levels in the survey.

The two versions of the survey instrument were the same, other than the scenario description. In the health survey, the researchers were asked to evaluate studies that were developed to value a public-health intervention, whereas in the environment survey the commodity was an environmental intervention. Each SP task presented two hypothetical SP studies. The studies were to be used by a government agency to inform a policy decision requiring monetary values for an intervention (referred to as the *recommendation question* from now on). Researchers were reminded that these are not necessarily state-of-the-art studies, but good-practice applications of SP methods. Subjects were presented with five different options: (1) recommend Study A only, (2) recommend Study B only,

[3] The scope test was defined as the 'test of whether subjects' values vary positively and commensurately with an increasing quantity of the good offered' in the survey instrument. We did not specify whether this is a within-subject or between-subjects test for the ABM question formats.
[4] The attribute levels were defined as 'Yes: The study confirmed that income is a statistically significant determinant of WTP. No: The study did not explore whether income is a statistically significant determinant of WTP'.

Table 2. Example choice task comparing study options

Feature		Study A	Study B
Question format		Open-ended	Dichotomous choice with uncertainty or other follow-up
Experimental design		*Not applicable*	Bid structure based on literature review
Survey development		Included pretest interviews	Included pretest interviews
Administration mode		Mail	Telephone
Analysis	Scope test	Pass test	Pass test
	Income a significant determinant of WTP	No	Yes
	Scenario rejection, outliers, consistency, 'don't know' responses	Adjustment	Adjustment
	Econometrics	Basic modelling (such as logit or probit)	Basic modelling (such as logit or probit)

Please indicate which study you would recommend	○ A only	○ B only	○ Both	○ Neither	○ Not Sure

Please indicate your evaluation:	○ A is better than B	○ B is better than A

(3) recommend both studies, (4) recommend neither study, and (5) 'not sure' what to recommend. A follow-up question asked subjects to choose which hypothetical study is better (referred to as the *relative-quality question* from now on). We asked the recommendation question before the relative-quality question to avoid anchoring the subjects on one alternative.[5] Table 2 provides an example of an SP question.

We employed a fractional factorial experimental design, which selects a particular sample of factorials that are capable of estimating the parameters (Hensher *et al*., 2005). We employed SAS macros to construct a D-efficient experimental design resulting in 24 pairs of study options (Huber and Zwerina, 1996; Kanninen, 2002; Kuhfeld, 2005). To reduce subject burden, the choice tasks were blocked into four sets of six choice questions. Each subject was randomly assigned to receive one of these four sets of questions plus an additional two method questions and two questions used to construct a transitivity test.

All subjects were presented with the same two method questions. The first method question varied only in the question-format and the experimental-design attributes, while keeping all other attributes constant (Appendix A). This question contrasts an ABM question format that uses a D-efficient design with a CVM question format with a follow-up question that uses a bid structure based on a pilot study. The second method question varies the attributes associated with the statistical analysis, other attributes held constant. In this question, Study A fails a scope test, finds that income is a significant determinant of WTP, adjusts for scenario rejection and outliers, and uses advanced modelling (Appendix B). Study B passes the scope test, finds that income is not a significant determinant of WTP, does not adjust for scenario rejection and outliers, and uses basic modelling.

The final two choice questions were designed to assess the validity of each subject's stated preferences with a transitivity test. Transitivity requires that if subjects indicate they prefer study A to study B at one point in the question sequence and indicate they prefer study B to study C at another point, then they should also prefer study A to study C in a third question.

Subjects also may adopt decision heuristics to simplify the choice tasks, such as selecting alternatives based on the best level of a single attribute rather than evaluating the advantages and disadvantages of all attributes. This heuristic can bias preference estimates if the choices do not reflect actual evaluations. We checked for this dominant-preference response pattern.

The survey also included standard demographic items (e.g. age, gender, education) as well as a number of items about subjects' research experience, such as professional affiliation, primary research interest, years of research experience, and number of publications.

[5] We would like to thank Dr. Joel Huber for his suggestion on this anchoring effect.

Sample

We identified 265 authors who have published valuation papers in health journals and 320 authors who have published valuation papers in environmental journals between 1995 and 2006. Subjects accessed the web-enabled survey via a link in an email. We were unable to obtain email information for some of these researchers and some email addresses were invalid. We sent the web-survey to 173 health researchers and 263 to environment researchers.[6] We received 91 completed surveys from health researchers and 136 completed surveys from environment researchers.

Model

Survey subjects were presented with a series of evaluation tasks involving choices between two study options. According to hedonic-utility principles, all alternatives over which subjects make choices (such as study options) can be thought of as being composed of a set of attributes. Each alternative i is described according to a vector of distinct attribute levels (X_i). Extending the familiar random-utility modelling framework to include professional judgements regarding relative-quality and policy-recommendation judgements, the random quality judgement associated with each alternative is assumed to be a function of these attributes plus a random error term

$$\begin{aligned} J_i &= Q_i + \varepsilon_i \\ Q_i &= X_i\beta \end{aligned} \tag{1}$$

where

- J_i is the random quality judgement;
- Q_i is the determinate part of the quality-assessment function for study i;
- X_i is a vector of attribute levels for study i;
- β is a vector of attribute parameters (marginal utilities); and
- ε_i is a random error.

In this article, the policy-recommendation and relative-quality judgements for a given study i are functions of coefficients for effects-coded categorical study attributes. Thus the empirical specification for judgement J_i is

$$\begin{aligned} J_i = &\left[\beta_{OpenEnded} + \beta_{Bounded} + \beta_{RankingRating} + \beta_{GradedPairs}\right] \\ &+ \left[\beta_{LiteratureReview} + \beta_{Defficient}\right] + \left[\beta_{Focus\_pretest} + \beta_{Pretest}\right] \\ &+ \left[\beta_{Mail} + \beta_{Phone} + \beta_{Face}\right] + \left[\beta_{PassScope} + \beta_{NoScope}\right] \\ &+ \beta_{Income\_significant} + \beta_{Adjusted\_outliers} + \beta_{Advanced\_statistics} \end{aligned}$$

For the policy-recommendation judgement questions, the dependent variable was whether or not subjects recommended the study for policy use in the recommendation question. If both studies were recommended, then the dependent variable was '1' for both studies, and if neither was recommended, then the dependent variable was '0' for both. 'Not sure' responses were dropped from the analysis. In the relative-quality judgement question, the dependent variable was '1' for the chosen study and '0' for the other study.

We used multivariate, random-parameters or mixed-logit regression to estimate judgement parameters for each attribute level. Mixed logit avoids potential estimation bias from unobserved taste heterogeneity in discrete-choice models by estimating a distribution of tastes for each parameter (Revelt and Train, 1998). In addition, because each subject provided responses to multiple choice questions, we estimated a mixed-logit panel model to account for within-subject correlation.

Some subjects did not answer some of the relative-quality questions where we forced them to choose which study is better. We estimated a tobit model to explain the observed number of times subjects skipped the relative-quality question (Greene, 2003). We hypothesize that skipping is a result of relative inexperience with and knowledge of the methods being evaluated, or the subjects did not think one study was better than the other.

III. Results

Survey population

Table 3 summarizes and compares researcher characteristics and their work experience from the two samples. The environment sample was significantly larger than the health sample. Compared to the health sample, the environment sample had more years of experience working on SP methods, more researchers who are knowledgeable about CVM, more SP publications, more researchers who served as an advisor to a graduate student, and more researchers who review manuscripts for publication ($p<0.05$ for all). Although the distribution of primary professional affiliation was significantly different ($p<0.01$), the number of academic researchers was not significantly different between the two groups. The number of researchers who are knowledgeable about ABM methods was higher in the environment sample at the 10% significance level. The number of researchers based in the US and the number of male researchers also were significantly higher in the environment sample ($p<0.05$).

The descriptive statistics overall indicate that the majority of researchers had an academic affiliation (68% for health, 79% for environment), used SP methods for both methodological and empirical research (47% for health, 56% for environment), served as an advisor to a government agency (64% for health, 74% for environment), spent on average 16 years in health or environment related work, and about one third of their work in valuation methods focused on CVM and one-third on ABM.

Internal validity tests

Four health researchers (about 4%) and three environment researchers (2%) failed the transitivity test. Two environmental researchers always picked the alternative that passed the scope test. We dropped the following observations from the

[6] We sent both survey versions to researchers who had published in both health and environmental journals.

Table 3. Summary of researcher characteristics

Characteristic	Health researchers ($N = 91$)	Environment researchers ($N = 136$)	p-value
Professional affiliation, N (%)			0.008
Academic	62 (68)	107 (79)	0.764
Government	1 (1)	12 (9)	
Nonprofit	5 (5)	7 (5)	
For-profit consulting	8 (9)	10 (7)	
Industry	5 (5)	0	
Primary research interest, N (%)			
Cost-effectiveness analysis	33 (36)	*NA*	
Cost-benefit analysis	36 (40)	*NA*	
Risk-benefit analysis	7 (8)	*NA*	
Pharmacoeconomics	22 (24)	*NA*	
Health services research	34 (37)	*NA*	
Population health	10 (11)	*NA*	
Health outcomes research	38 (42)	*NA*	
Environmental policy analysis	*NA*	97 (71)	
Food and agricultural marketing	*NA*	11 (8)	
Rural development	*NA*	11 (8)	
Land use management	*NA*	40 (29)	
Natural resource management	*NA*	51 (38)	
Environmental policy analysis	*NA*	97 (71)	
Demand assessment	8 (9)	33 (24)	
Impact evaluation	6 (7)	29 (21)	
Primary focus of SP research, N (%)			0.587
Methodological studies	10 (11)	22 (16)	
Empirical or applied studies	27 (30)	38 (28)	
Both methodological and empirical/applied studies	43 (47)	76 (56)	
Years in health or environment related research, mean (SD)	16 (7.5)	16 (8.4)	0.793
Years working on SP methods, mean (SD)	11 (7.4)	14 (8)	0.013
Served as an advisor to a government agency, N (%)	58 (64)	100 (74)	0.758
How knowledgeable on CVM, N (%)			<0.001
Not knowledgeable	6 (7)	0	
A little knowledgeable	10 (11)	7 (5)	
Somewhat knowledgeable	30 (33)	27 (20)	
Quite knowledgeable	19 (21)	57 (42)	
Very knowledgeable	16 (18)	45 (33)	
How knowledgeable on ABM, N (%)			0.053
Not knowledgeable	4 (4)	5 (4)	
A little knowledgeable	16 (18)	12 (9)	
Somewhat knowledgeable	18 (20)	48 (35)	
Quite knowledgeable	22 (24)	39 (29)	
Very knowledgeable	12 (13)	32 (24)	
Number of peer-reviewed publications, N (%)			0.089
1–10 publications	11 (12)	30 (22)	
11–20 publications	13 (14)	31 (23)	
21–50 publications	20 (22)	44 (32)	
More than 50 publications	28 (31)	31 (23)	
Number of peer-reviewed publications on SP methods, N (%)			0.009
1–10 publications	56 (62)	85 (63)	
11–20 publications	9 (10)	23 (17)	
21–50 publications	4 (4)	22 (16)	
More than 50 publications	0	5 (4)	
Served as advisor to a graduate student, N (%)	37 (41)	99 (73)	0.001
Review manuscripts for publication, N (%)	59 (65)	126 (93)	0.011
Distribution of valuation work in last 2 years, mean (SD)			
CVM	27% (34)	32% (30)	0.345
ABM	38% (34)	44% (33)	0.272
Statistical modelling, N (%)			
OLS	19 (21)	28 (21)	0.274
Binary logit or probit	19 (21)	43 (32)	0.521
Multinomial logit or probit	20 (22)	58 (43)	0.049
Conditional logit	11 (12)	42 (31)	0.019

(*continued*)

Table 3. Continued

Characteristic	Health researchers ($N=91$)	Environment researchers ($N=136$)	p-value
Nested logit	5 (5)	19 (14)	0.151
Random-parameters or mixed logit	14 (15)	54 (40)	0.004
Ordered probit	11 (12)	11 (8)	0.089
Bivariate-probit model	4 (4)	11 (8)	0.542
Nonparametric	11 (12)	20 (15)	0.833
Latent class	2 (2)	10 (7)	0.195
Fixed-effects panel	4 (4)	11 (8)	0.542
Random-effects panel	20 (22)	17 (13)	0.004
Hierarchical Bayes	2 (2)	9 (7)	0.259
WTP space	2 (2)	6 (4)	0.590
PhD highest degree, N (%)	62 (68)	118 (87)	0.537
Economics degree	41 (45)	101 (74)	
Country of residence, N (%)			
US	30 (33)	77 (57)	0.040
UK	12 (13)	17 (13)	0.409
Europe	15 (16)	16 (12)	0.218
Australia	8 (9)	7 (5)	0.114
Canada	4 (4)	3 (2)	0.203
Gender, N (%)			0.028
Male	44 (48)	103 (76)	
Age, mean (SD)	45 (10)	47 (8.8)	0.371

analysis: (1) Two researchers (one from each sample) who had no variation in his or her answers to the choice questions. (2) Two *responses* that recommended Study B (A) only in the recommendation questions, but then reported that Study A (B) was better than Study B (A). The answers to the relative-quality and recommendation questions are expected to be correlated but there may be other reasons for this choice behaviour. Since there were only two, we dropped these responses (not researchers) from the data.

Methodological questions

In the first method question, where we compared an ABM study and a CVM study, 18% of the health researchers recommended the former, 10% recommended the later study, and 48% recommended both studies. Among the environmental researchers, 14% recommended the former, 6% recommended the later study, and 59% recommended both studies. The proportions were statistically significantly different for both the recommendation question ($p=0.05$) and the relative-quality question ($p=0.01$). In the relative-quality question, about 54% of the health researchers selected the ABM format, while 23% selected the CVM format. Among the environmental researchers, 44% and 34% selected the ABM format and the CVM format, respectively.

The second method question varied the analysis attributes. About 63% of the health researchers and 48% of the environmental researchers recommended both studies ($p<0.01$) in the recommendation question. The majority of the researchers in both samples selected Study A which fails the scope test but all other analysis attributes are better.

Parameters estimates

Attribute levels were effects coded. The parameter for the omitted category is the negative sum of the included-category parameters (Hensher *et al*., 2005). Thus zero is the mean effect for each attribute, and positive and negative coefficients are interpreted relative to the mean effect. Likelihood-ratio tests rejected pooling of the data from the two samples ($p < 0.01$).

Relative-quality model

Table 4 contains the effects-coded, mixed-logit coefficients and SDs and their SEs for relative quality. To facilitate comparisons and avoid problems of confounding taste and scale in judgement parameters, we rescaled the parameter estimates from 0 to 10, where 0 corresponds to the worst level and 10 corresponds to the best level (Swait and Louviere, 1993). Figure 1 presents the scaled estimates. Open-ended was the least-favoured question format for CVM studies ($p<0.01$). While the choice format was the favourite question type for an ABM survey, environmental researchers made no significant distinction between graded pairs and ratings/rankings ($p=0.30$).[7] A CVM bid design based on a pilot study was favoured over a bid design based on literature review, and

[7] We also investigated whether the choice of question format systematically differed if a response format was compared within its own method (CVM versus CVM, or ABM versus ABM) versus it was compared with a question format from the other method (CVM versus ABM). The frequency of researchers' choosing a CVM study was not statistically significantly different when CVM was one of the alternatives against an ABM response format versus CVM was the response format in both alternatives (p=0.302). Similarly, the frequency of researchers' choosing an ABM study was not statistically significantly different when ABM was one of the alternatives against a CVM response format versus ABM was the response format in both alternatives (p=0.955).

Table 4. Effects-coded mixed-logit parameter estimates and their standard errors for the relative-quality questions

	Health sample				Environment sample			
	Coefficient	SE	SD	SE	Coefficient	SE	SD	SE
Question format								
CVM open ended	−7.71	2.34	4.67	1.25	−2.04	0.42	1.16	0.50
CVM single or double	3.49	2.49	4.78	1.22	1.07	0.32	0.66	0.41
CVM with follow-up	4.22	1.55			0.97	0.31		
ABM ranking/rating	−4.61	1.20	0.94	0.57	−0.68	0.27	0.26	0.40
ABM graded pairs	0.46	0.68	6.79	1.77	−0.42	0.26	0.23	0.31
ABM choice	4.14	1.12			1.10	0.25		
Experimental design								
CVM literature review	−1.86	0.84	7.26	1.67	−0.53	0.29	1.03	0.26
CVM pilot	1.86	0.84			0.53	0.29		
ABM D-efficient design	4.47	1.25	4.76	1.16	0.40	0.21	0.50	0.25
ABM catalog based	−4.47	1.25			−0.40	0.21		
Survey development								
Focus groups and pretest	5.11	1.24	2.22	0.58	1.11	0.21	0.75	0.21
Only focus groups	2.79	0.84	4.10	1.05	0.36	0.16	0.81	0.24
None	−7.89	1.89			−1.47	0.26		
Administration mode								
Mail	−0.44	0.54	4.10	1.05	−0.06	0.22	0.77	0.28
Phone	−0.55	0.84	6.35	1.52	−0.02	0.21	0.01	0.23
Face	2.54	0.83	6.12	1.87	0.54	0.25	0.58	0.42
Web	−1.54	0.73			−0.47	0.23		
Scope test								
Pass test	0.20	0.48	3.57	0.98	0.46	0.19	0.57	0.16
No test	0.81	0.66	2.45	0.67	0.56	0.25	0.58	0.22
Fail test	−1.01	0.44			−1.02	0.21		
Income/WTP								
Income significant	1.32	0.79	1.66	0.55	0.33	0.12	0.73	0.16
Income not	−1.32	0.79			−0.33	0.12		
Scenario rejection								
Adjustment	4.48	1.11	2.61	0.76	0.52	0.13	0.59	0.13
No adjustment	−4.48	1.11			−0.52	0.13		
Statistics								
Advanced	0.54	0.91	0.49	0.44	0.97	0.48	0.10	0.22
Basic	−0.54	0.91			−0.97	0.48		
Interactions*								
Know CVM_advanced	−2.66	0.97	0.28	0.33	−1.58	0.51	0.52	0.14
Know CVM_basic	2.66	0.97			1.58	0.51		
Know ABM_advanced	4.96	1.34	3.99	1.06	0.26	0.31	0.43	0.19
Know ABM_basic	−4.96	1.34			−0.26	0.31		
Number of researchers	65				117			
Number of observations	626				1,127			
Log-likelihood	−280.0899				−551.4233			

Notes: The omitted categories do not have the SD estimates and the SEs of the SDs.
*Know CVM_advanced is an interaction variable between being knowledgeable about CVM and advanced modeling. Know CVM_basic is an interaction variable between being knowledgeable about CVM and basic modeling. Know ABM_advanced is an interaction variable between being knowledgeable about ABM and advanced modeling. Know ABM_basic is an interaction variable between being knowledgeable about ABM and basic modeling.

a D-efficient design was favoured over a catalog-based design in surveys ($p<0.01$). Researchers clearly favoured focus groups and/or pre-test of the survey instrument. The added value of pretests to focus groups was significant for environmental researchers ($p=0.00$) but it was not for health researchers ($p=0.22$). Health researchers significantly prefer in-person interviews to other administration modes ($p<0.01$), but no particular mode was favoured in the environment sample. Interestingly, *absence* of a scope test was not significantly different than *passing* a scope test, while both were significantly favoured over *failing* the test. As expected, researchers favoured income as a significant determinant of WTP, and adjustments should be made for scenario rejection, outliers, inconsistency, and don't know responses. While, advanced modelling was somewhat favoured over basic modelling, this difference was not significant in the main-effects model.

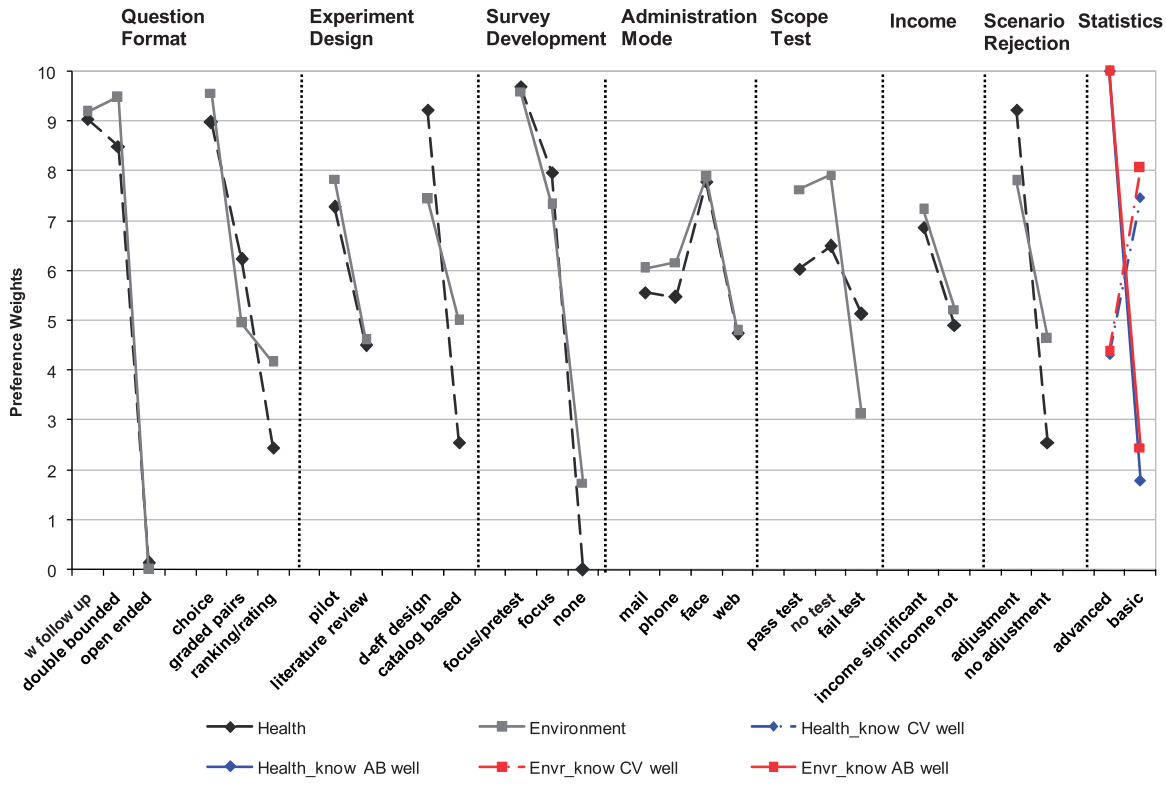

Fig. 1. Scaled preference estimates for the relative-quality questions
Note: 'Know CVM well' and 'Know AB well' represent the interaction variables on statistics in the relative-quality question.

We conducted *t*-test comparisons of parameter estimates between the two groups after adjusting for the scale differences between the two data sets. D-efficient designs were favoured significantly more by the health sample than the environment sample at the 10% significance level ($p = 0.08$). Environmental researchers valued *passing* a scope test ($p = 0.06$) and disvalued *failing* a scope test ($p = 0.00$) significantly stronger than the health researchers. The penalization of scenario rejection was stronger by health researchers than environmental researchers ($p = 0.07$). We tried interactions between the question format and the other attributes. However, this model did not improve the specification since coefficients on most interaction variables were insignificant. We also tried interaction variables with researcher characteristics, such as years of experience, gender, number of publications, and how knowledgeable researchers are about the method. Only the interactions with knowledge of methods and statistical modelling variables were significant in both samples. Researchers who reported they were knowledgeable about ABM methods favoured advanced modelling over basic modelling. This difference was not significant for researchers who were knowledgeable about CVM.

The parameters with the largest SDs for the mixed-logit parameter estimates, indicating the least consensus among researchers, were the literature-review bid design for a CVM study, graded-pairs for an ABM study, telephone and in-person interviews for the health sample, and open-ended CVM question and bid design based on literature review for the environment sample. The parameters with the smallest SDs, indicating the greatest consensus among researchers in both samples, were the ranking/rating question format and advanced modelling.

To facilitate comparisons among the attributes in our design, the relative importance of each attribute was measured as the percent contribution to the difference between the best and worst level for each attribute to the quality assessment. Figure 2 presents the relative importance of each attribute within the levels used in this study. According to the health sample, survey development, question format and statistics were the most-important attributes. The three most-important attributes were the same for the environment sample, although the order of importance was different. Also, scope test was the most important 4th attribute for the environment sample, whereas it was the least important for the health sample.

Recommendation model

Table 5 and Fig. 3 present the parameters for the recommendation model. We dropped the observations with 'not sure' responses to the recommendation question. If someone recommended both studies, we assumed each study satisfied the recommendation threshold separately.

The only statistically significant differences between the two samples were for experimental design ($p = 0.04$) and survey development ($p = 0.03$). A bid design based on a pilot study was favoured over a design based on literature review by

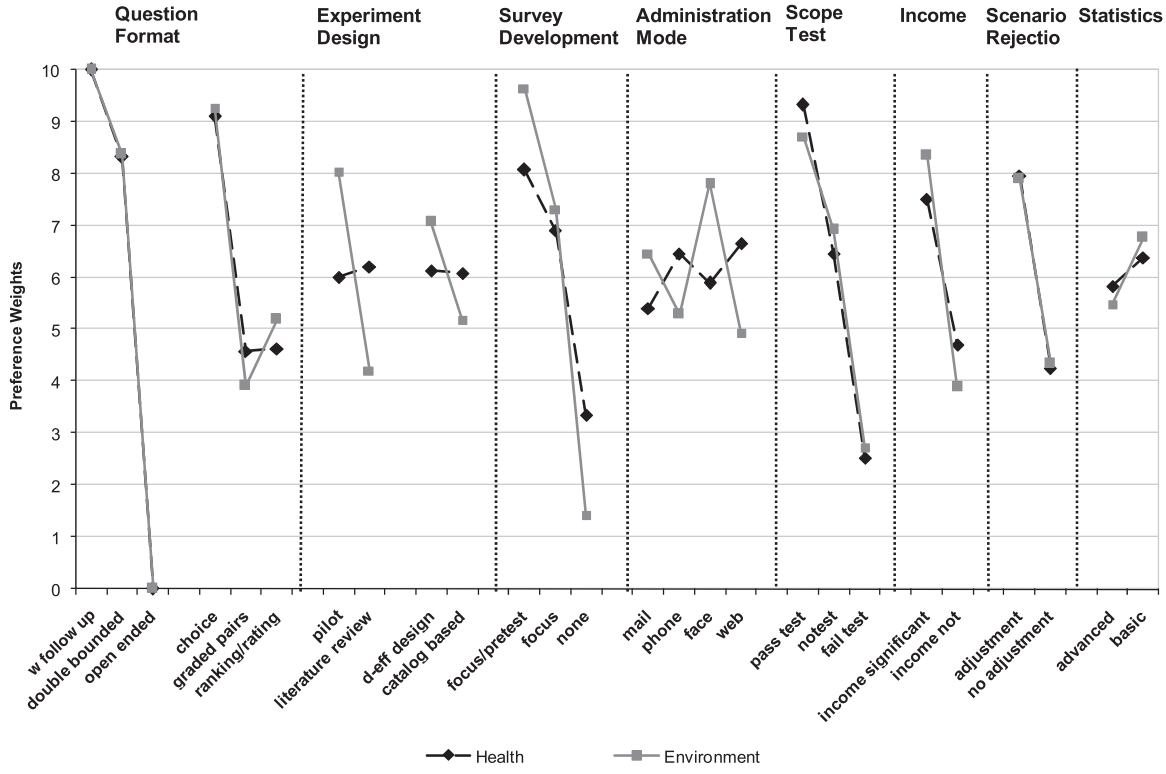

Fig. 2. Scaled preference estimates for the recommendation questions

environmental researchers, whereas the type of the experimental design did not matter for health researchers when recommending a study for a policy application. Although 'no focus group studies or pretests' was the least favoured in both samples, environmental researchers penalized this more than the health researchers.

We also compared the judgements in the relative-quality and recommendation models. Health researchers had significantly different judgements for several attributes in the two models ($p < 0.05$ for all), whereas environmental researchers did not. For health researchers, the type of experimental design was important in the relative-quality model, but it did not matter for policy recommendations. The importance of conducting focus groups and pretest interviews and the importance of adjustments to scenario rejection diminished in the recommendation model. Scope test became a very important attribute for policy recommendation, moving in the opposite direction compared to changes in the other attributes.

Item-nonresponse analysis

The researchers skipped (did not answer) 1.8 (out of 10) questions on average, and 45% of the sample skipped at least one relative-quality question. The *t*-test comparisons showed that the mean number of questions skipped was not statistically significantly different between the health and environment samples ($p = 0.336$); and it was not statistically significantly different for a question that compares a CVM and an ABM question format ($p = 0.499$). Table 6 presents the tobit estimates for the pooled data from both samples where the dependent variable is the number of times a researcher skipped a relative-quality question. Researchers tended to skip more questions if they spent less time taking the survey, suggesting that nonresponse partly was a time-saving strategy. The number of item nonresponses significantly increased as the number of 'not sure' responses increased in the recommendation questions. Academic affiliation, years in SP research, being a technical government advisor, number of publications, reviewing SP studies, being knowledgeable about ABM or CVM methods, being a health researcher, and answering a question that compares a CVM and an ABM question format did not have significant effects on nonresponse.

IV. Discussion

To our knowledge, this is the first study to use SP methods to elicit judgements about good-practice methods. Our results have several important implications. First, the parameter estimates suggest that the single or double-bounded CVM format, the dichotomous choice CVM format with a follow-up question, and the ABM choice format are the favourite question types. However, when researchers were asked to compare these two question formats directly in the first method question, most chose the ABM choice format over the single or double-bounded CVM format. These results suggest that researchers, especially health researchers, favoured ABM choice-format studies over other ABM and CVM approaches.

Table 5. Effects-coded mixed-logit parameter estimates and their standard errors for the recommendation questions

	Health sample				Environment sample			
	Coefficient	SE	SD	SE	Coefficient	SE	SD	SE
Question format								
CVM open ended	−1.20	0.23	0.83	0.27	−1.03	0.16	0.69	0.21
CVM single or double	0.44	0.18	0.02	0.28	0.38	0.14	0.05	0.43
CVM with follow-up	0.77	0.21			0.65	0.16		
ABM ranking/rating	−0.29	0.19	0.70	0.16	−0.16	0.13	0.29	0.19
ABM graded pairs	−0.30	0.15	0.00	0.21	−0.37	0.14	0.51	0.15
ABM choice	0.59	0.17			0.53	0.12		
Experimental design								
CVM literature review	0.02	0.15	0.53	0.20	−0.32	0.11	0.28	0.24
CVM pilot	−0.02	0.15			0.32	0.11		
ABM D-efficient design	0.01	0.11	0.19	0.17	0.16	0.09	0.16	0.13
ABM catalog based	−0.01	0.11			−0.16	0.09		
Survey development								
Focus groups and pretest	0.39	0.13	0.43	0.16	0.59	0.11	0.40	0.10
Only focus groups	0.16	0.09	0.05	0.23	0.20	0.08	0.39	0.16
None	−0.54	0.13			−0.79	0.11		
Administration mode								
Mail	−0.14	0.16	0.38	0.14	0.06	0.12	0.39	0.16
Phone	0.07	0.15	0.15	0.16	−0.14	0.12	0.41	0.14
Face	−0.04	0.16	0.25	0.21	0.28	0.12	0.05	0.17
Web	0.11	0.14			−0.20	0.11		
Scope test								
Pass test	0.63	0.14	0.66	0.13	0.44	0.08	0.05	0.17
No test	0.07	0.12	0.06	0.20	0.14	0.09	0.31	0.12
Fail test	−0.71	0.14			−0.57	0.09		
Income/WTP								
Income significant	0.28	0.08	0.02	0.10	0.38	0.07	0.34	0.08
Income not	−0.28	0.08			−0.38	0.07		
Scenario rejection								
Adjustment	0.37	0.08	0.17	0.22	0.30	0.07	0.40	0.10
No adjustment	−0.37	0.08			−0.30	0.07		
Statistics								
Advanced	−0.05	0.09	0.24	0.14	−0.11	0.07	0.35	0.10
Basic	0.05	0.09			0.11	0.07		
Number of researchers	65				117			
Number of observations	660				1195			
Log-likelihood	−773.1346				−1358.3278			

Note: The omitted categories do not have the SD estimates and the SEs of the SDs.

Second, the main statistically significant differences between the two samples were scope-test and survey-development attributes. Environmental researchers penalized failing a scope test and conducting no focus groups much more strongly than health researchers. Scope test had a much higher importance rank in the environment sample than the health sample. As one of the reviewers suggested the importance of the scope test among environmental economists might have arisen from the history of scope tests and high profile valuation cases in the environmental area.

Third, the result regarding statistical modelling is particularly interesting. The main-effects, relative-quality questions and recommendation questions suggest that advanced modelling was not judged to be better than basic modelling. This may be a result of the type of the scenario we asked researchers to evaluate. Statistical modelling seems not to be important for a study that is used in a policy application. Judgements might have been quite different if we asked researchers to evaluate state-of-the-art studies, not best-practice studies. Nevertheless, it is interesting that researchers did not seem to be particularly concerned about study quality although the results may influence actual policies. Possible explanations include a belief that simple statistical models produce results sufficiently close to advanced statistical models or a belief that simple models are easier to explain to decision makers and thus more likely to influence decisions. In pretest interviews, some experienced researchers expressed the view that advanced statistical analysis could overcome weaknesses in other features of the study. Thus another possible explanation may be that if you develop and administer an SP survey carefully, then it is less likely that you will need to use advanced modelling to deal with data problems. However, the

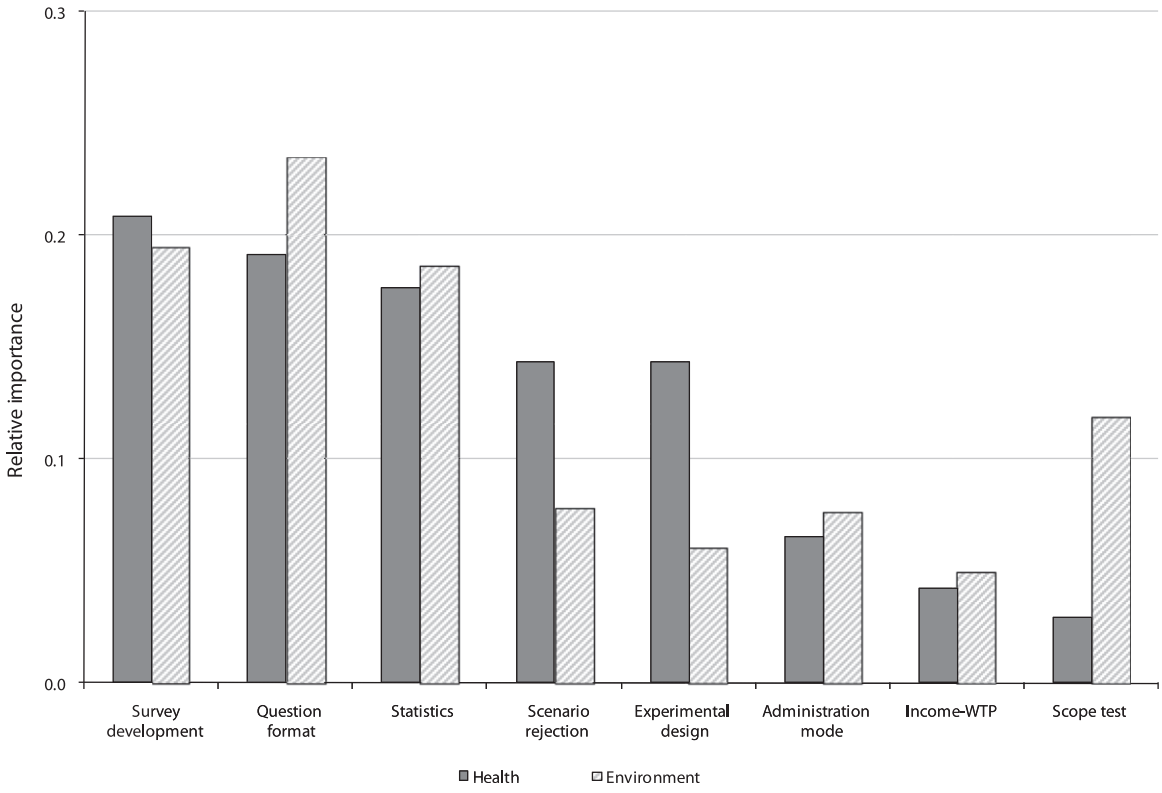

Fig. 3. Relative importance of each attribute

interaction models showed that researchers who have experience with ABM methods favoured advanced modelling over basic modelling. This was the only attribute for which experience with methods had a significant influence.

Finally, we found that health researchers did not discriminate much among study features for policy recommendations, but they had stronger judgements over different attribute levels when deciding which study is better. This result implies that although health researchers had specific judgements on which study features were better, their quality judgements were not very consistent with their judgements about the acceptability of studies for policy analysis. On the other hand, environmental researchers had similar judgements over the study attributes for the two types of questions.

These results are subject to several limitations and qualifications. One inherent limitation is that the choice tasks asked subjects to evaluate hypothetical scenarios. Although we defined a scenario as recommending studies for general policy purposes, the evaluation could be different depending on the specific policy involved. Second, we unfortunately could not include other study attributes, such as sample size, budget and sampling method because of practical limitations. For the same reasons, we also had to limit the number of levels we used for each attribute. We controlled for such omissions by asking researchers to assume everything not included in the design is the same for all studies shown. Third, we assumed that researchers apply the same methodological standards in applying these methods in different applications within the same field (for example in a study on asthma versus multiple sclerosis within the health field). However, some of the characteristics of the study, especially the response format, could be sensitive to the specific context. For example, evaluation of a good with multiple attributes versus a single attribute could lead researchers to choose a different question format. Fourth, it is possible that researchers could have behaved strategically and chose the alternatives that mimic their favourite approach or the format they usually use in their own studies.[8] Competition for publication and other forms of professional recognition linked to research methods could be a powerful incentive for conscious or unconscious strategic responses. This kind of strategic behaviour may happen to make their favourite approach to be recommended for policy analysis. Fifth, the relatively high level of nonresponse could be related to the fact that the relative-quality question forced researchers to choose one of the alternatives. Although we were aware of the possible consequences of this type of response format, we used a forced-response format because subjects are more likely to choose the opt-out or status-quo alternative when the variability in the alternatives is small (Dhar, 1997), which was likely to be the case in this study. While we provided every possible alternative in the recommendation question, the relative-quality question was designed to obtain as much trade-off information among the attributes as possible, at the possible expense of some level of

[8] We would like to thank one of the referees for suggesting the possibility of strategic behaviour.

Table 6. Tobit model of nonresponse frequency in the relative-quality question

Researcher characteristics	Coefficient	SE	*p*-value
Academic affiliation	−0.318	0.873	0.716
Years on SP methods	0.200	0.177	0.261
Square of years on SP methods	−0.005	0.005	0.268
Technical advisor	−1.311	0.870	0.134
Number of published SP studies	0.106	0.100	0.290
Square of number of published SP studies	−0.001	0.002	0.424
Review SP manuscripts	0.001	1.261	1.000
Knowledgeable about CVM	−0.106	1.352	0.938
Knowledgeable about ABM	−0.993	1.050	0.346
Survey time	−0.226	0.068	0.001
Square of survey time	0.003	0.001	0.001
Health sample	0.045	0.795	0.955
Number of skipping a question that compares a CVM and an ABM question format	−0.243	0.278	0.382
Number of times a researcher was 'not sure' which alternative to recommend in the recommendation questions	0.739	0.132	0.000
Constant	4.228	3.185	0.186

Notes: The dependent variable is the number of times a subject skips the relative-quality question.
Pseudo $R^2 = 0.076$ and Log-likelihood = −318.94.

nonresponse. The tobit regression results confirmed that researchers tend to skip more frequently in the relative-quality question if they were 'not sure' which alternative to recommend in the recommendation questions.

Judgements about relative quality may not be particularly informative about publishing standards. While our results identify some study features that are important for policy applications, standards appropriate for publishing an SP study in a scientific journal may be different. Future studies may investigate researcher judgements for developing state-of-the-art studies, or investigate how stated judgements from this study relate to past and current revealed preferences based on journal publications and to current standards on state-of-the-art application of SP studies in policy analysis in the health and environmental fields. In addition, the focus of this study was WTP estimation. Monetary valuation of benefits is far more acceptable in environmental economics than in health economics. The different experience in applications may have influenced comparisons between the two groups apart from the methods themselves.

This study is the first attempt to quantify the degree of consensus among researchers about SP methods. Our results suggest that consensus about good-practice SP methods varies considerably among study features and among researchers with different amounts and kinds of research experience. This study may help health researchers, who have adapted environmental-valuation methods for health applications, to assess how successful and complete the technology transfer between environment and health applications has been.

SP methods have been mandated for regulatory analysis in environmental policy, transportation, food safety, and other applications for as much as 25 years in the US and other countries. However, this is not the case for health. The lack of a regulatory mandate for health valuation means incentives are weaker than in other areas of applied economics for training younger health researchers in best-practice methods and for methods development in health applications. Our study suggests that environmental researchers have similar criteria for evaluating quality of a study and recommending a study for policy use. However, health researchers do not seem to apply the strict criteria they had for relative quality to policy recommendations. This may raise a question about whether the analytical standards for health-policy studies are sufficient.

In summary, the findings from this study provide insights on researchers' views on good-practice methods in SP studies. Survey development, question format and statistics were identified as the most important attributes when designing SP studies. However, for policy relevance researchers judged conducting focus groups and pretest interviews, utilizing a CVM dichotomous-choice with certainty or other follow-up question format or a choice ABM format and passing a scope test as the most important factors.

References

Bateman, I., Carson, R., Day, B., Hanemann, W. M., Hanley, N., Hett, T., Jones-Lee, M., Loomes, G., Mourato, S., Özdemiroglu, E., Pearce, D. W., Sugden, R. and Swanson, J. (2002) *Economic Valuation with Stated Preference Techniques: A Manual*, Edward Elgar, Cheltenham.

Bennett, J. and Adamowicz, V. (2001) Some fundamentals of environmental choice modeling, in *The Choice Modeling Approach to Environmental Valuation* (Eds) J. Bennett and R. Blamey, Edward Elgar, Cheltenham, pp. 37–69.

Boyle, K. (2003) Contingent valuation in practice, in *A Primer on Nonmarket Valuation* (Eds) P. Champ, K. Boyle and T. Brown, Kluwer Academic Publishers, Norwell, pp. 111–70.

Bryan, S., Gold, L., Sheldon, R. and Buxton, M. (2000) Preference measurement using conjoint methods: an empirical investigation of reliability, *Health Economics*, **9**, 385–95.

Carson, R., Hanemann, M., Kopp, R., Krosnick, J., Mitchell, R., Presser, S., Ruud, P., Smith, V. K., Conaway, M. and Martin K. (1998) Referendum design and contingent valuation: the NOAA panel's no-vote recommendation, *The Review of Economics and Statistics*, **80**, 335–8.

Carlsson, F. and Matinsson, P. (2001) Do hypothetical and actual marginal willingness to pay differ in choice experiments? Application to the valuation of environment, *Journal of Environmental Economics and Management*, **41**, 179–92.

Dhar, R. (1997) Consumer preference for a no-choice option, *Journal of Consumer Research*, **24**, 215–31.

Diener, A., O'Brien, B. and Gafni, A. (1998) Health care contingent valuation studies: a review and classification of the literature, *Health Economics*, **7**, 313–26.

Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. J. and Stoddart, G. L. (2005) *Methods for the Economic Evaluation of Health Care Programmes*, Oxford University Press, USA.

Green, P. E. and Rao, V. R. (1971) Conjoint measurement for quantifying judgmental data, *Journal of Marketing Research*, **8**, 355–63.

Green, P. E. and Wind, Y. (1975) New way to measure consumers' judgments, *Harvard Business Review*, **53**, 107–17.

Greene, W. H. (2003) *Econometric Analysis*, Prentice Hall, Upper Saddle River, New Jersey.

Hanley, N., Ryan, M. and Wright, R. (2003) Estimating the Monetary Values Of Health Care: Lessons From Environmental Economics, *Health Economics*, **12**, 3–16.

Hensher, D. A., Rose, J. and Greene, W. H. (2005) *Applied Choice Analysis: A Primer*, Cambridge University Press, Cambridge, England.

Holmes, T. and Adamowicz, V. (2003) Attribute-based methods, in *A Primer on Nonmarket Valuation* (Eds) P. Champ, K. Boyle and T. Brown, Kluwer Academic Publishers, Norwell, pp. 171–220.

Huber, J. and Zwerina, K. (1996) The importance of utility balance in efficient choice designs, *Journal of Marketing Research*, **33**, 307–17.

Johnson, F. R. and Desvousges, W. H. (1997) Estimating stated preferences with rated-pair data: environmental, health, and employment effects of energy programs, *Journal of Environmental Economics and Management*, **34**, 79–99.

Johnson, F. R., Fries, E. and Banzhaf, H. S. (1998) Valuing morbidity: an integration of the willingness-to-pay and health-status index literatures, *Journal of Health Economics*, **16**, 641–65.

Kanninen, B. J. (2002) Optimal design for multinomial choice experiments, *Journal of Marketing Research*, **39**, 214–27.

Kuhfeld, W. (2005) *Marketing Research Methods in SAS: Experimental Design. Choice, Conjoint, and Graphical Techniques*, SAS Institute Inc., Cary, NC, USA. Available at http://support.Sas.com/techsup/technote/ts722.Pdf (accessed 8 February 2012).

Lancsar, E. and Louviere, J. (2006) Deleting 'irrational' responses from discrete choice experiments: a case of investigating or imposing preferences?, *Health Economics*, **15**, 797–811.

Lloyd, A. (2003) Threats to the estimation of benefit: are preference elicitation methods accurate?, *Health Economics*, **12**, 393–402.

McIntosh, E., Clarke, P. M., Frew, E. J. and Louviere, J. J. (2010) *Applied Methods of Cost-benefit Analysis in Health Care*, Oxford University Press, Oxford.

McIntosh, E. and Ryan, M. (2002) Using discrete choice experiments to derive welfare estimates for the provision of elective surgery: implications of discontinuous preferences, *Journal of Economic Psychology*, **23**, 367–82.

Miguel, F. S., Ryan, M. and Amaya-Amaya, M. (2005) 'Irrational' stated preferences: a quantitative and qualitative investigation, *Health Economics*, **14**, 307–22.

Miguel, F. S., Ryan, M. and Scott, A. (2002) Are preferences stable? The case of health care, *Journal of Economic Behavior and Organization*, **48**, 1–14.

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

Office of Management and Budget (OMB) (2003) OMB draft guidelines for the conduct of regulatory analysis and the format of accounting statements, *Federal Register*, **68**, 5519.

Olsen, J. and Smith, R. (2001) Theory versus practice: a review of willingness-to-pay in health and health care, *Health Economics*, **10**, 39–52.

Randall, A., Ives, B. and Eastman, C. (1974) Bidding games for evaluation of aesthetic environmental improvements, *Journal of Environmental Economics and Management*, **1**, 132–49.

Revelt, D. and Train, K. (1998) Mixed logit with repeated choices: households' choices of appliance efficiency level, *Review of Economics and Statistics*, **80**, 647–57.

Ryan, M. and Amaya-Amaya, M. (2004) 'Threats' to and hopes for estimating benefits, *Health Economics*, **14**, 609–19.

Ryan, M. and Bate, A. (2001) Testing the assumptions of rationality, continuity and symmetry when applying discrete choice experiments in health care, *Applied Economic Letters*, **8**, 59–63.

Ryan, M. and Miguel, F. S. (2003) Revisiting the axiom of completeness in health care, *Health Economics*, **12**, 295–307.

Ryan, M., Scott, D., Reeves, C., Bate, A., Van Teijlingen, E., Russell, E., Napper, M. and Robb, C. M. (2001) Eliciting public preferences for healthcare: a systematic review of techniques. *Health Technology Assessment (Winchester, England)*, **5**, 1–186.

Schwappach, D. L. B. and Strasmann, T. J. (2005) Quick and dirty numbers? The reliability of a stated-preference technique for the measurement of preferences for resource allocation, *Journal of Health Economics*, **25**, 432–48.

Shiell, A., Seymour, J., Hawe, P. and Cameron, S. (2000) Are preferences over health states complete?, *Health Economics*, **9**, 47–55.

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

Appendix A: Methodological Question 1

Feature		Study A	Study B
Question format		Choice format	Dichotomous choice with uncertainty or other follow-up
Experimental design		D-efficient design	Bid structure based on a pilot study
Survey development		Included pretest interviews	Included pretest interviews
Administration mode		Web-based	Web-based
Analysis	Scope test	Pass test	Pass test
	Income a significant determinant of WTP	Yes	Yes
	Scenario rejection, outliers, consistency, 'don't know' responses	Adjustment	Adjustment
	Econometrics	Advanced modelling (such as nonparametric modelling and mixed logit)	Advanced modelling (such as nonparametric modelling and mixed logit)

Please indicate which study you would recommend	○	○	○	○	○
	A only	B only	Both	Neither	Not Sure

Please indicate your evaluation:	○	○
	A is better than B	B is better than A

Appendix B: Methodological Question 2

Feature		Study A	Study B
Question format		Choice format	Choice format
Experimental design		Catalog-based design	Catalog-based design
Survey development		Included focus groups, pretest interviews, and reviewed by technical experts	Included focus groups, pretest interviews, and reviewed by technical experts
Administration mode		Web-based	Web-based
Analysis	Scope test	Fail test	Pass test
	Income a significant determinant of WTP	Yes	No
	Scenario rejection, outliers, consistency, 'don't know' responses	Adjustment	No adjustment
	Econometrics	Advanced modelling (such as nonparametric modelling and mixed logit)	Basic modelling (such as logit or probit)

Please indicate which study you would recommend	○	○	○	○	○
	A only	B only	Both	Neither	Not Sure

Please indicate your evaluation:	○	○
	A is better than B	B is better than A