University of Zurich

Socioeconomic Institute
Sozialökonomisches Institut

Working Paper No. 0313

Validity of Discrete-Choice Experiments Evidence for Health Risk Reduction

Harry Telser and Peter Zweifel

October 2003, revised February 2005

Socioeconomic Institute
University of Zurich

Working Paper No. 0313

Validity of Discrete-Choice Experiments Evidence for Health Risk Reduction

October 2003, revised February 2005

Author's addresses | Harry Telser
E-mail: telser@soi.unizh.ch

Peter Zweifel
E-mail: pzweifel@soi.unizh.ch

Publisher | Sozialökonomisches Institut
Bibliothek (Working Paper)
Rämistrasse 71
CH-8006 Zürich
Phone: +41-1-634 21 37
Fax: +41-1-634 49 82
URL: www.soi.unizh.ch
E-mail: soilib@soi.unizh.ch

Validity of Discrete-Choice Experiments Evidence for Health Risk Reduction*

Harry Telser [a,b] and Peter Zweifel [a]

[a] *Socioeconomic Institute, University of Zurich, Switzerland*

[b] *Plaut Economics, Regensdorf, Switzerland*

October 2003, revised: February 2005

Correspondence to:

Peter Zweifel

Socioeconomic Institute

University of Zurich

Hottingerstr.10

CH-8032 Zurich

Switzerland

Tel.: +41-1-634-3720

Fax: +41-1-634-4987

e-mail: pzweifel@soi.unizh.ch

Abstract

There is growing interest in discrete-choice experiments (DCE) as a method to elicit consumers' preferences in the health care sector. Increasingly this method is used to determine willingness-to-pay (WTP) for health-related goods. However, its external validity in the health care domain has not been investigated until today. This paper examines the external validity of DCE concerning the reduction of a health risk. Convergent validity is examined by comparing the value of a statistical life with other preference elicitation techniques, such as revealed preference. Criterion validity is shown by comparing WTP values derived from stated choices in the experiment with those derived from actual choices made by the same individuals. Both tests provide strong evidence in favor of external validity of the DCE method.

JEL Classification Numbers: C25, C52, D12, I18, I19

Keywords: Discrete-Choice Experiments (DCE), Willingness-to-Pay (WTP), Validity, Risk Reduction, Hip Protectors.

---

* Financial support by the Swiss Council for Accident Prevention (CAP) is gratefully acknowledged. Also, authors wouold like to thank two anonymous referees for their perceptive comments and criticisms.

1 Introduction

Discrete-choice experiments (DCE) are designed to allow individuals to express their preferences for non-marketed goods or goods which do not exist yet. They have become very popular in transportation economics (Hensher, 1997) and environmental economics (Adamowicz und Boxall, 2001). Recently, there have been several applications of DCE in the health domain (for a survey see Ryan and Gerard, 2003). However, being hypothetical, DCE are subject to the criticism that they may fail to be related to (and ultimately predict) actual choices. Outside health, Louviere and Woodsworth (1983) have presented evidence suggesting that DCE can be used to explain modal choice in transportation, while Ghosh (1986) has found that they contribute to explaining the choice of a shopping mall. DCE with regard to health have so far been hardly related to actual choices (Ryan, 2004; Ryan and Gerard, 2003).

The present contribution seeks to fill this gap, and in a context that may be deemed particularly difficult because one of the product attributes distinguished is risk reduction. It is well known that probabilities and changes in probabilities pose particular problems in surveys and experiments. The specific product in question is a hip protector that serves to all but eliminate the risk of fracture of the femur in aged persons. With an estimated mortality of up to 20 percent (Dubey *et al.*, 1998) and treatment cost around US$ 30,000 per case and year (Barefield, 1996), fractures of the femur are far from trivial. Therefore, their prevention is of some economic interest; on the other hand, hip protectors must be worn to be effective. This study relates willingness-to-pay (WTP) values estimated in a DCE (the stated choice) to the same respondents' later willingness to participate in a wearing trial of two months' duration (the actual choice). In this way, this study is likely to be the first in the health domain to confront results from a DCE with actual choices.

In section 2, the issues of reliability and validity as quality attributes of DCE are discussed in a review of the literature. In section 3, the methods used in the present DCE for establishing relevant product attributes and estimating marginal willingness-to-pay (MWTP) are described. Results with regard to convergent and criterion validity of DCEs are presented in section 4. Section 5 contains a discussion and conclusion.

2 Reliability and validity of DCE results

In the course of a DCE, respondents are confronted with a sequence of product varieties with changing product attributes. In each case, they are asked to indicate which of the presented varieties they prefer. In this way an indifference hyperplane in attribute space is approximated. Since one of the attributes is out-of-pocket price, the marginal rate of substitution between a particular attribute and price indicates the marginal willingness-to-pay (MWTP) for that attribute. The WTP for the product as a whole can be determined by integrating the MWTP values up to the corresponding attribute levels and aggregating (Johnson and Desvousges, 1997).

Following Jöreskog and Goldberger (1975) (see also Schoenberg and Arminger, 1989), let y_1 and y_2 denote two measurements of a latent variable X. Since the determinants of X (neglected in eq. (1) below for simplicity) may change between the two observations, X in general will have unobserved values X_1 and X_2. However, measurements may be contaminated by an irrelevant latent variable Z. In the present context, y_1 and y_2 are 'observed' (calculated) WTP values that depend on a latent quantity X (such as marginal rate of substitution) that may differ between subsamples ($X_1 \neq X_2$); however, they by hypothesis should not be a function of e.g. political attitudes Z. Then, observations are generated according to the measurement model

$$y_1 = \lambda_1 X_1 + \mu_1 Z_1 + e_1, \quad y_2 = \lambda_2 X_2 + \mu_2 Z_2 + e_2, \tag{1}$$

with (λ_1, μ_1, λ_2, μ_2) denoting the loadings of measurements on latent variables and (e_1, e_2) stochastic i.i.d. error terms. There are several sources of systematic error. Relevant product attributes may not have been recognized (the X vector is too short to begin with), seemingly irrelevant attributes may have been excluded (the X vector has been erroneously shortened), or the underlying indirect utility function may have been wrongly specified (the structural model determining X is wrong). Random errors are always present; in DCE, they are even part of the specification, which is based on the random utility model.

2.1 Reliability

Reliability requires that measurements can be reproduced at least on average, which means that loadings should be constant ($\lambda_1 = \lambda_2$, $\mu_1 = \mu_2$) and random errors zero on expectation, $E(e_1) = E(e_2) = 0$. There are several ways to check for reliability (see e.g. Litwin, 1995): The test-retest method benefits from repeated measurement; in that case, eq. (1) applies directly. Parallel testing involves the simultaneous use of two slightly different instruments; in this case, y_1 and y_2 are two different indicators, which should have loadings λ_1 and λ_2, that differ in a predictable way provided $\mu_1 = \mu_2 =$ const. Finally, in the alternate-form method the sample is split, with part of the observations reserved for reestimation using a variant of the measurement method. Here, y_1 and y_2 refer to the two segments of the sample that again should induce loadings λ_1 and λ_2 that deviate in a predictable way.

DCE applied to health have but rarely been tested for reliability (Ryan and Gerard, 2003). A likely reason is that additional observations come at a steep marginal cost, causing sample size to be rather small to begin with. Nevertheless, Farrar and Ryan (1999) using the alternate-form method found that the sequence of presentation of attributes did not change the setting of priorities in hospitals. Bryan *et al.* (2000) used the test-retest procedure by asking the same persons about their preferences for treatment of knee injuries within two weeks. The authors found a high degree of conformity between the two DCE. Telser and Zweifel (2002) using the same data as presented here performed an out-of-sample test in a DCE dealing with fractures of the femur. They concluded that DCE are an instrument of high reliability for measuring WTP for the reduction of a health risk.

2.2 Validity

The validity requirement is more stringent, requiring not only $\lambda_1 = \lambda_2$ indicators, but also $\mu_1 = \mu_2 = 0$ in eq. (1) ($\lambda_1 \neq \lambda_2$ or $\mu_1 \neq \mu_2 \neq 0$ only if the maintained hypothesis makes a prediction regarding the relative quality of the two indicators). Thus, not only must measurement be reliable but also free of systematic, uncontrolled bias that could be caused by a variable Z that is irrelevant by hypothesis. There are several types of validity (see again Litwin, 1995). Economists often distinguish between internal and external validity. *Internal validity* refers to (1) the rigor with which the study was conducted (e.g., the study's design, the care taken in conducting the interviews, and decisions concerning what was and what was not measured); it makes certain values of λ_1 and λ_2 credible; and (2) the extent to which the designers of a study have taken into account alternative explanations for any causal relationships they explore, for example by explicitly testing for an influence of Z. *External validity* refers to the generalizability of the results to other populations, settings, and circumstances; thus, it makes credible that $\lambda_1 = \lambda_2$ if the two values are estimated from two arbitrary samples or populations. In this paper we focus on two variants of *external* validity, viz. convergent and criterion validity.

Convergent validity obtains if different methods that are designed to generate information about the same theoretical construct X have convergent results (y_1, y_2). In the present case, true WTP is the theoretical construct, which can be elicited e.g. by contingent valuation (CV), revealed preference (RP), and DCE methods. If DCE has convergent validity, the WTP values obtained should be similar to those elicited with the other methods, provided these benchmark methods constitute valid elicitation techniques; thus $\lambda_1 = \lambda_2$ although y_1 and y_2 are generated by different methods. *Criterion validity* obtains if the results of a method correspond with an external criterion. For example, WTP calculated in a DCE can be compared to WTP observed in actual choices, which provide the external criterion; in this case, an alternative with a known value of λ is available.

DCE for environmental goods have been tested for convergence validity. Thus, Mason and Quigley (1983) found that hedonic regression (where market prices are related to product attributes) resulted in estimates that were similar to those derived from DCE. The estimates of both methods were close to the 'true' values of a Monte Carlo simulation. A similar conclusion was drawn by Gegax and Stanley (1997) who compared WTP values for safety obtained from hedonic regression and DCE. They found no significant differences between the hedonic and DCE estimates. More recent results from environmental economics confirm convergence validity between RP and DCE (see e.g. Hensher *et al.*, 1999; Louviere *et al.*, 1999). However, juxtaposing CV and DCE results, Boxall *et al.* (1996) arrived at WTP estimates that were twenty times higher in the case of CV. At the same time, these authors were able to show that DCE results were in conformity with RP, i.e. with actual choice behavior. Mark and Swait (2004) come to the same conclusion with regard to physicians' evaluation of an alcohol medication and actual prescribing decisions.Still other studies cast doubt on the validity of the CV method (see, e.g., Hausman, 1993; Nocera *et al.*, 2002) or suggest superiority of DCE over CV (Ryan, 2004)

Outside health, the criterion validity of DCE has been tested on several occasions, causing Louviere (1988) to conclude that DCE constitute a valid instrument for explaining and predicting individual behavior on actual markets. More recently Carlsson and Martinsson (2001) observed that measured WTP for a public good did not differ much between a DCE with hypothetical payments and a subsequent DCE with real payments. However, in a control group that was only confronted with actual payments, WTP was substantially lower (Johansson-Stenman and Svedsäter, 2003). While this latter finding is not compatible with criterion validity, it may not carry over to private goods, to which a majority of health care services belong. In the context of health, the survey by Ryan and Gerard (2003) shows the validity of DCE to be still an open issue; specifically convergence and criterion validity seem not yet to have been tested.

Conclusion 1: In contrast to environmental economics, discrete-choice experiments concerning health-related goods have been rarely checked for reliability and hardly for external validity, the latter involving convergent and criterion validity.

3 Methods

For the present study, aged persons were interviewed in a pretest in order to establish the relevant product attributes of a hip protector. Following standard DCE procedures, the product attributes of such a hip protector were preliminarily assumed to be protective effect, wearing comfort, ease of handling, change of appearance, and out-of-pocket cost. In the pretest ($N = 17$), it turned out that the dimension 'appearance' was judged unimportant by a clear majority, justifying its exclusion from the main survey (see Telser and Zweifel, 2002 for details). The importance of the product attributes was again ascertained in the field survey. A majority of respondents judged all of the retained four attributes to be very important, notably the attribute 'out-of-pocket cost'. This may be surprising in health care systems with comprehensive health insurance coverage. However, in Switzerland insureds are subject to a degree of cost sharing especially in ambulatory care. This also makes a decision situation involving out-of-pocket payment less hypothetical for respondents. The levels of the attributes were defined as follows (see Table 1). 'Protective effect' (PROT), symbolizing the risk reduction from an unknown individual level, takes on the values of 100, 75, and 50%. The choice of levels reflects the high effectiveness of existing variants of hip protectors (see e.g., Lauritzen et al., 1993; or Ekman et al., 1997). 'Ease of handling' (HAND) varies between 3 (very easy to put on) and 1. The same holds for 'wearing comfort' (COMF). The 'out-of-pocket cost' (COST) ranges from CHF 0 (which reflects the scenario of health insurance fully covering the hip protector in the future) to CHF 200 (US$ 133 at 1998 exchange rates), bracketing the price (CHF 80) typically paid by institutional purchasers. The reference scenario is the status quo (no purchase of hip protector). For a sample choice scenario, see the Appendix.

Table 1 Product attributes and levels retained in the main survey

Attributes	Label	Levels	Value labels
Protective effect	PROT	100% protection from hip fracture	100
		75% protection from hip fracture	75
		50% protection from hip fracture	50
Ease of handling	HAND	Handling is easy	3
		Handling is somewhat complicated	2
		Handling is very complicated	1
Wearing comfort	COMF	Comfortable to wear	3
		Somewhat uncomfortable to wear	2
		Uncomfortable to wear	1
Out-of-pocket cost	COST	CHF 0 (US$ 0)	0
		CHF 75 (US$ 50)	75
		CHF 150 (US$ 100)	150
		CHF 200 (US$ 133)	200

Since the first three attributes have 3 levels each, while COST has 4, the number of possible scenarios amounts to a total of 108 (= 3*3*3*4). Techniques have been developed to reduce the number of possible scenarios while still being able to infer utilities for all combinations of levels of the attributes (see Louviere et al., 2000). Using the ORTHOPLAN procedure programmed in the software package SPSS, the design was reduced to 23 scenarios ensuring orhogonality w.r.t. main effects. These 23 variants were split into two (randomly presented) subsets featuring a different sequence of presentation of the hip protectors to avoid boredom and bias on the part of respondents. With regard to each variant, respondents had to indicate whether or not they would buy the product.

The main survey consisted of 522 face-to-face interviews (of about 45 min length on average) with individuals aged 70 and older in the Summer of 1998. The sample is representative with regard to age and sex of the independently living Swiss subpopulation. At the end of the interview, participants in the survey were asked whether they were willing to take part in a wearing test of a specific hip protector (HIPS®) free of charge (COST = 0) which was carried out three months later. Before the trial, participants had to rate the product in terms of PROT, HAND, and COMF of HIPS®. The intentions stated in the last part of the interview can be interpreted as an actual choice because participants were expected to document their trial in a diary during three months. This provides the basis for the test of criterion validity of DCE presented in section 4.2.

4 Results

4.1 Probit estimation

The theoretical basis for estimation is provided by the random utility model (McFadden, 1974) which assumes that the difference in the systematic component of utility must exceed the difference in its stochastic component for the individual to opt for the alternative rather than the status quo. Since the error term is assumed to be normally distributed, the difference is normally distributed as well, calling for Probit estimation. Since PROT, HAND, and COMF are categorically defined, a Wald test is used to determine whether their linear representation is admissible (for details, see Telser and Zweifel, 2002). Linearity need not to be rejected. Finally, since the same individuals made several choices, the data set has a panel structure. To reflect this fact, a random effects specification was chosen for the error term. Thus, $e_{ij} = \eta_{ij} + \upsilon_i$, with υ_i denoting the individual-speficic component and η_{ij} the general, component of the error term that may also vary across scenarios.

The estimation result is shown in Table 2 (= Table 4 in Telser and Zweifel, 2002). All product attributes retained are highly significant and have the expected signs. This provides some evidence in favor of criterion validity because attribute selection was based on an independent pretest. In addition, although protective effect (PROT) was defined in terms of probability, a concept many individuals are not very familiar with, it has a strongly significant positive effect on the propensity to opt for the alternative. The COST attribute has a negative and highly significant coefficient, a result that implies that inferences on WTP can safely be made.

Table 2 Random effects probit estimates

Variable	Coefficient	Std. err.	z	P > |z|
Protective effect (PROT)	0.0141***	0.0013	11.1	0.000
Ease of handling (HAND)	0.3325***	0.0313	10.6	0.000
Wearing comfort (COMF)	0.6628***	0.0361	18.4	0.000
Out-of-pocket cost (COST)	-0.0029***	0.0003	-10.1	0.000
CONSTANT	-3. 9812***	0.1868	-21.3	0.000

Number of obs.	3714	Deviance	3269. 53
Chi^2 (15)	569.31	Dispersion	0.8815
Prob > chi^2	0.000	ρ	0.1879

The coefficients shown in Table 2 can be used to calculate WTP values (details in Telser and Zweifel, 2002). For instance, the marginal WTP for 1 percentage point of risk reduction amounts to an estimated US$2.49 (s.e. 0.50) for individuals with a monthly income of less than US$2000 but US$5.00 (s.e.1.55) for individuals with US$2000 and more – a difference that is in keeping with microeconomic theory provided that protection against risk is a normal good (Zweifel and Breyer 1997, ch. 2).

4.2 Convergent validity: value of a statistical life

Willingness-to-pay for a marginal risk reduction can be extrapolated to obtain the value of a statistical life. For example, the estimates cited at the end of the preceding paragraph would suggest that to low-income individuals in Switzerland, eliminating the risk of breaking the femur entirely is worth US$249 and to those with monthly incomes above US$2000, about US$500, among individuals aged 70 and over. Breaking the hip being lethal with a certain probability, values of a statistical life can be deduced from these figures (see below). Since CV as well as RP methods have been used for estimating the value of life, comparing these values creates an opportunity for testing convergent validity. Focus will be on the RP alternative, which can be considered the 'gold standard' from the point of view of economics; in addition, it has been used more frequently than CV. RP estimates are most often based on wage differentials associated with differences in health risks between occupations (Viscusi, 1992). There is an abundance of such estimates; therefore, the entries of Table 3 are derived from four surveys.

The results of the present study must be made comparable with those shown in Table 3. First, calculated WTP is for a reduction of the risk of breaking the femur rather than the risk of death. In advance of the choice experiment, it was ascertained that respondents did understand the risks associated with a fractured femur. Before making their hypothetical decisions in the DCE, respondents were informed about their risk of breaking the femur as well as the resulting risk of death. Mortality rates given fracture of the femur were used according to age class to estimate the implied relative reduction in mortality due to this particular cause (Hubacher and Ewert, 1997; Lippuner *et al.*, 1997). The associated marginal WTP values were then integrated for a protective effect of 100 percent which allowed the computation of the value of a statistical life, amounting to 1.9 mn. US$ for individuals aged 70 –75.

Second, the values cited in Table 3 are based on individuals around forty years of age. On theoretical grounds, Shepard and Zeckhauser (1982) predict that the relationship between the value of a statistical life and age is inversely U-shaped, with a maximum value around the age of forty. This prediction has been empirically confirmed by several authors (Carthy *et al.*, 1999; Jones-Lee *et al.*, 1985; Mount *et al.*, 2000). The estimate of the present study is made comparable with the international evidence as follows. Its mean value of life of US$ 1.9 mn, pertaining to individuals in the 70-75 age bracket, is adjusted to an age of 40 years using the empirically found minimum and maximum values of the difference between ages 40 and 70 of the studies cited. The calculated minimum value amounts to US$ 2.4 mn, the maximum value to US$ 4.0 mn.

Table 3 Value of a statistical life, in million US$ (1998 values) [a)]

Survey	Preferred estimate [b)]	Min	Max
Viscusi (1992) [c)]	5.1	0.75	24.1
U.S. Environmental Protection Agency (1999)	6.0	0.75	16.9
de Blaeij *et al.* (2000)	2.7	0.15	15.6
Sommer *et al.* (1999)	1.4	0.65	5.7
This study	3.2	2.4	4.0

[a)]. Value around age 40, see text for details

[b)] Value which is preferred by the authors of the survey article

[c)] Two studies with extremely low values are not taken into account.

The preferred estimate of US$ 3.2 mn is the average of these two values. It is somewhat lower than those from US studies but in keeping with the meta analysis by de Blaeij *et al.* (2000), where European values generally fall short of their US counterparts. As for Switzerland, there is one study (Baranzini and Ferro Luzzi, 2001) that is based on RP methods. It comes up with a value of statistical life between US$ 6 and 10 mn, which is on the high side when compared with the European values given in Sommer *et al.* (1999).

Conclusion 2: The age-adjusted WTP results from the present discrete-choice experiment study are comparable with other estimates based on revealed preference methods, providing preliminary evidence in favor of their convergent validity.

An additional piece of evidence comes from *divergent validity*, which is the logical counterpart of convergent validity, requiring that values of a statistical life should differ if they are based on different concepts. Specifically, the human capital approach neglects aspects such as pain and suffering that enter the determination of WTP, a difference that may well be of particular importance among individuals of an advanced age. Therefore, the human capital approach should result in lower estimates of the value of a statistical life than the WTP approach adopted in this study. A study by Buzby *et al.* (1996) indeed finds a value of US$ 58,000 (1998 values) for 79-year old individuals (the average age in the present sample), which lies far below the US$ 670,000 (at that age) calculated from the results of the present study.

4.3 Criterion validity: actual and stated choice

Whereas in the preceding section, focus was on estimating MWTP for risk reduction, the objective here is to calculate WTP for the specific hip protector HIPS® presented to respondents at the end of the experiment. This allows to relate the decision to participate in the wearing trial (which is an indicator of actual WTP) to the stated WTP value, amounting to a test of criterion validity. Ideally, one would attempt to predict participation in the trial based on estimated individual WTP. However, these values derive from utility differences, causing socioeconomic characteristics to drop out from the estimating equation unless they enter through interactions with product attributes, in particular the cost attribute, permitting variation of the marginal utility of income (Johnson and Desvousges, 1997). The present sample is limited to individuals aged 70 and more and living independently, resultingin a high degree of homogeneity in terms of measured socioeconomic characteristics. This fact may explain why these interaction terms did not attain statistical significance, precluding calculation of individual WTP values. Even if these values could have been calculated, it is doubtful whether predicting the participation decision would have been successful because only 16 percent of the sample (83 respondents out of 522) chose to participate in the trial. This makes estimation of a distribution function difficult.

In this situation, prediction has to be replaced by postdiction. The maintained hypothesis is that WTP for participants is higher than that of nonparticipants. Therefore, differences in WTP values must be calculated for such postdiction,. There are three sources of differences: (1) Participants in the trial may rate the product attributes of HIPS® higher; (2) They may put a higher MWTP on these attributes; and (3) their WTP for the entire product as the combination of attributes may be higher. With regard to (1) attribute levels, note that the protective effect (PROT) of HIPS® is set to 100 percent in view of biomechanical studies showing that wearing this hip protector completely eliminates the risk of fracture of the femur (Denoth, 1998). Similarly, out-of-pocket cost (COST) is equal to zero because the protector was offered free of charge for the trial. As to the two remaining attributes, Table 4 suggests that participants indeed rate both ease of handling and wearing comfort of the protector HIPS® higher than do nonparticipants; however, the difference is not statistically significant.

Table 4 Rating of product attributes of HIPS® by participants and nonparticipants before the wearing trial

	Participants		Nonparticipants	
Product attributes	Mean rating	Std.dev. [b]	Mean rating	Std.dev. [b]
Protective effect (PROT) [a]	100	-	100	-
Ease of handling (HAND)	2.51	0.67	2.19	0.75
Wearing comfort (COMF)	2.36	0.63	2.02	0.74
Out-of-pocket cost (COST) [a]	0	-	0	-

[a] Objective value

[b] Standard deviation calculated conservatively using bootstrapping (1,000 repetitions); two alternative methods (Delta, Fieller) resulted in smaller values.

With regard to (2) MWTP values for product attributes, the only improvement over the status quo is protective effect, whereas even the highest levels of ease of handling and wearing comfort still imply a deterioration compared to the status quo. Postdiction therefore amounts to saying that the marginal valuation of this one advantage relative to the two disadvantages should be higher in the participant group. In Table 5, MWTP for risk reduction (PROT) is slightly higher among participants in the trial. In addition, they exhibit significantly lower MWTP for the other two attributes than nonparticipants. While this is puzzling, it does imply that participants are willing to sacrifice more ease of handling and wearing comfort for a given amount of risk reduction than nonparticipants, a difference in marginal rate of substitution that accords with expectations.

Table 5 Marginal WTP for product attributes of HIPS® among participants and nonparticipants, in US$

	Participants		Nonparticipants	
Product attributes [a]	Marginal WTP	Std.dev. [b]	Marginal WTP	Std.dev. [b]
Protective effect (PROT)	3.35	0.71	3.21	0.59
Ease of handling (HAND)	56.96	15.25	91.68	16.38
Wearing comfort (COMF)	122.93	23.59	162.53	23.51
Correction term	-685.84	68.54	-1009.95	55.07

[a] Measured per percentage point in the case of PROT and per unit in the case of HAND and COMF.

[b] Standard deviation calculated conservatively using bootstrapping (1,000 repetitions); two alternative methods (Delta, Fieller) resulted in smaller values.

When it comes to (3) calculating mean WTP for HIPS®, MWTP values need to be integrated up to the relevant attribute levels and aggregated over attributes (Johnson and Desvousges, 1997). The higher attribute ratings of participants (taken from Table 4) combine with correction terms to overcompensate their lower MWTP for HAND and COMF. Correction terms reflecting the different constant terms need to be introduced; they amount to the ratio between the estimated constant and the coefficient of the cost attribute from the respective random effects probit specification (see Table 2). A negative correction term can be interpreted as the reservation price that has to be paid to compensate for the utility loss of having to wear a hip protector at all. Such a reservation price mirrors an unfavorable basic attitude towards the product in question and is often called status quo bias. One would therefore expect the correction term (if negative) to be less marked among participants in the trial than nonparticipants. The entries in the bottom row of Table 5 again confirm this expectation.

Calculated in this way, WTP values for HIPS® are displayed in Table 6. They are to be tested against zero for the following reason. Participation choices permit to determine upper and lower bounds of actual WTP values. For participants the lower bound must be the out-of-pocket cost (which is zero), while this constitutes the upper bound for nonparticipants. From this, one can infer that mean stated WTP should be greater than zero for participants but less than zerofor nonparticipants. This expectation is borne out (see Table 6) in that participants exhibit a mean value of stated WTP that is significantly positive whereas nonparticipants in the wearing trial display a significantly negative WTP on average.

One might argue that a higher WTP among participants was to be expected in view of Table 5; however, the evidence presented in Table 6 goes beyond this. Indeed, it is the postulated indirect utility function that yields the prediction that participants in the wearing trial derive more utility from nonprice product attributes than disutility from price. Since price is set to zero, their WTP should be positive and nonparticipants' WTP negative (reflecting their negative utility derived from nonprice attributes). Since they confirm the utility framework underlying DCE, the results presented in Table 6 thus constitute rather strong evidence in favor of DCE having criterion validity.

Table 6 Willingness-to-pay for HIPS® for participants and nonparticipants, in US$

	Participants	Nonparticipants
Willingness-to-pay for HIPS®	82.54	-160.07
Standard deviation [a)]	12.78	30.69
T-value (against 0)	6.46	-5.22

[a)] Standard deviation calculated conservatively using bootstrapping (1,000 repetitions); two alternative methods (Delta, Fieller) resulted in smaller values.

Conclusion 3: Differences between participants and nonparticipants in the wearing trial with regard to mean rating of product attributes and marginal WTP values for product attributes tend to conform with theoretical expectations. Moreover, differences in stated total WTP correspond to predicted differences in actual WTP. Taken together, these results strongly suggest criterion validity of the discrete-choice experiments undertaken.

As a qualification, it is appropriate to point to the possibility of respondents' individual characteristics (that were not included in the regression for reasons stated at the beginning of section 4.3) influencing both their stated WTP and the decision to participate in the wearing trial, resulting in a spuriously close correspondence between the two. Therefore, Conclusion 3 should not be interpreted as claiming a causality running from DCE responses to actual choice.

5 Summary and conclusion

Measuring willingness-to-pay (WTP) for health-related goods is of considerable importance because the financiers of health – be they governments or health insurers – can use this information for matching the provision of health care more closely with the preferences of consumers, thus enhancing rationality of decision making in this domain (Zweifel, 2001). However, this requires that measurements of WTP are valid in that they adequately reflect consumers' utility. A new and rapidly spreading method is the discrete-choice experiment (DCE), where respondents are confronted with product variants differing in attribute levels and asked to select the preferred one. Since one of these attributes is cost to the consumer, marginal WTP for an attribute can be derived from stated choices. By integrating marginal WTP to appropiate levels and aggregating across attributes, total WTP can be calculated as well. For all of its attractive features however, one important aspect of DCE has hardly been examined to date, viz. its external validity (Conclusion 1).

The present contribution purports to fill this gap by presenting evidence concerning convergent and criterion validity of a DCE involving a hip protector that could reduce the risk of fracture of the femur in an elderly Swiss population. Convergent validity requires the results of one method to come close to results based on an alternative that is considered valid. In this study, the stated WTP for risk reduction was used to derive the value of a statistical life, which can be juxtaposed to values based on established alternatives, in particular revealed preference. While lower than US estimates, the DCE-based value of a statistical life is close to European counterparts, suggesting convergent validity. Conversely, a comparison with estimates derived from the human capital method yields a higher value for the DCE alternative, as was to be expected for theoretical reasons (Conclusion 2).

Criterion validity provides a more powerful test in that it requires stated choices to be related to actual choices. The distinguishing feature of the present study is that respondents not only made hypothetical choices but also indicated their willingness to participate in a wearing trial, which amounts to an actual choice. Ideally, individual WTP values based on DCE should serve as predictors of actual choice. However, individual WTP values could not be calculated in this sample due to its high degree of homogeneity. This still leaves scope for postdiction in that participants should exhibit stated marginal and total WTP values that differ from those of nonparticipants. Indeed, among participants the attribute 'risk reduction' turns out to be more valuable in comparison to ease of handling and wearing comfort than among nonparticipants. As to stated total WTP, it should have the price of the product as an upper bound among nonparticipants and as a lower bound among participants in the wearing trial. Since price is zero (the trial being offered free of charge), nonparticipants are expected to exhibit negative and participants positive stated WTP. This expectation is confirmed as well, providing rather strong evidence in favor of DCE having criterion validity (Conclusion 3).

In all, the tests performed suggest that DCE may have a high degree of external validity even in difficult applications. For in this study, respondents were 79 years old on average, and they were asked to value risk reductions, i.e. changes in probability, a concept many people have difficulty with. On the other hand, at least two qualifications are in order. First, it was not possible to use stated WTP values as predictors of individual choices, which often constitutes the ultimate objective of such an exercise (Hall *et al.*, 2002). Second, in Swiss health care insureds are still used to a degree of cost sharing, which serves to make a decision situation involving out-of-pocket cost less hypothetical than in countries with comprehensive coverage or a national health service. Therefore, the final verdict about the merits of DCE is not out yet when it comes to truly public goods in health care.

6 References

ADAMOWICZ V, BOXALL P. Future Directions of Stated Choice Methods for Environment Valuation, Paper presented at the conference *Choice Experiments: A New Approach to Environmental Valuation*, London, 2001.

BARANZINI A, FERRO LUZZI G. The Economic Value of Risks to Life: Evidence from the Swiss Labour Market, *Schweizerische Zeitschrift für Volkswirtschaft und Statistik / Swiss Journal of Economics and Statistics* 2001, **137**, 149-70.

BAREFIELD E. Osteoporosis-Related Hip Fractures Cost $13 Billion to $18 Billion Yearly, *Food Review* 1996, **19**, 31-36.

BOXALL PC, ADAMOWICZ WL, SWAIT JD, *et al.* A Comparison of Stated Preference Methods for Environmental Valuation, *Ecological Economics* 1996, **18**, 243-53.

BRYAN S, GOLD L, SHELDON RJ, *et al.* Preference Measurement Using Conjoint Methods: An Empirical Investigation of Reliability, *Health Economics* 2000, **9**, 385-95.

BUZBY JC, ROBERTS T, JORDAN LIN CT, *et al. Bacterial Foodborne Disease: Medical Costs and Productivity Losses*, Economic Research Service, U.S. Department of Agriculture, Agricultural Economics Report Nr. 741, 1996.

CARLSSON F, MARTINSSON P. Do Hypothetical and Actual Marginal Willingness to Pay Differ in Choice Experiments? Application to the Valuation of the Environment, *Journal of Environmental Economics and Management*, 2001, **41**, 179-92.

CARTHY T, CHILTON S, COOKSON R, *et al. The Valuation of Benefits of Health and Safety Control: Final Report to the Health and Safety Executive*, London: Health and Safety Executive, 1999.

DE BLAEIJ A, FLORAX RJG, RIETVELD P, *et al. The Value of Statistical Life in Road Safety: A Meta-Analysis*, Working Paper TI 2000-089/3, Tinbergen Institute, Amsterdam, 2000.

DENOTH J. *Hüftprotektoren. Forschungsbericht zuhanden der Schweizerischen Beratungsstelle für Unfallverhütung bfu (Hip Protectors. Research Report for the Swiss Council for Accident Prevention CAP)*, Bern: Beratungsstelle für Unfallverhütung (bfu), 1998.

DUBEY A,. KOVAL KJ,. ZUCKERMAN JD. Hip Fracture Prevention: A Review, *American Journal of Orthopedics* 1998, **27**, 407-12.

EKMAN A, MALMIN H, MICHAËLSSON K. External Hip Protectors to Prevent Osteoporotic Hip Fractures, *Lancet*, 1997, **350**, 563-64.

FARRAR S, RYAN M. Response-Ordering Effects: A Methodological Issue in Conjoint Analysis, *Health Economics* 1999, **8**, 75-79.

GEGAX D, STANLEY LR. Validating Conjoint and Hedonic Preference Measures: Evidence from Valuing Reductions in Risk, *Quarterly Journal of Business and Economics* 1997, **36**, 31-54.

GHOSH A. The Value of a Mall and Other Insights from a Revised Central Place Model, *Journal of Retailing* 1986, **62**, 79-96.

HALL J, KENNY P, MADELEINE K, *et al.* Using Stated Preference Discrete Choice Modelling to Evaluate the Introduction of Varicella Vaccination, *Health Economics* 2002, **11**, 457-65.

HAUSMAN JA (ED.). *Contingent Valuation – a Critical Assessment*, Amsterdam, London, New York, Tokyo: North-Holland, 1993.

HENSHER DA. Stated Preference Analysis of Travel Choices: The State of Practice, in: *Transport Economics: Selected Readings*, Oum TH, Dodgson JS, Hensher DA *et al.* (eds.), Amsterdam: Harwood Academic in cooperation with the Korea Research Foundation for the 21st Century 1997, 81-109.

HENSHER DA, LOUVIERE JJ, SWAIT JD. Combining Sources of Preference Data, *Journal of Econometrics* 1999, **89**, 197-221.

HUBACHER M, EWERT U. *Das Unfallgeschehen bei Senioren ab 65 Jahren (Accidents Among Seniors Beyond the Age of 65)*, Bern: bfu, Report Nr. 32, 1997.

JOHANSSON-STENMAN O., SVEDSÄTER H, *Self-Image and Choice Experiments: Hypothetical and Actual Willingness to Pay*, Working Paper No 94, University of Göteborg, 2003.

JOHNSON FR,. DESVOUSGES WH. Estimating Stated Preferences with Rated-Pair Data: Environmental, Health, and Employment Effects of Energy Programs, *Journal of Environmental Economics and Management* 1997, **34**, 79-99.

JONES-LEE MW, HAMMERTON M,. PHILIPS PR. The Value of Safety: Results of a National Survey, *Economic Journal* 1985, **95**, 49-72.

JÖRESKOG K, GOLDBERGER AS. Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable, *Journal of the American Statistical Association* 1975, **70**, 631-39.

LAURITZEN JB, PETERSEN MM, LUND B. Effect of External Hip Protectors on Hip Fractures, *Lancet*, 1993, **341**, 11-13.

LIPPUNER K, VON OVERBECK J, PERRELET R, *et al.* Incidence and Direct Medical Costs of Hospitalizations Due to Osteoporotic Fractures in Switzerland, *Osteoporosis International*, 1997, **7**, 414-25.

LITWIN MS. *How to Measure Survey Reliability and Validity*, Thousand Oaks, London, New Delhi: SAGE Publications, 1995.

LOUVIERE JJ. Conjoint Analysis Modelling of Stated Preferences. A Review of Theory, Methods, Recent Developments and External Validity, *Journal of Transport Economics and Policy* 1988, **22**, 93-119.

LOUVIERE JJ, HENSHER DA, SWAIT JD. *Stated Choice Methods – Analysis and Application*, Cambridge: Cambridge University Press, 2000.

LOUVIERE JJ, MEYER RJ, BUNCH DS, *et al.* Combining Sources of Preference Data for Modeling Complex Decision Processes, *Marketing Letters* 1999, **10**, 205-17.

LOUVIERE JJ, WOODSWORTH G. Design and Analysis of Simulated Choice or Allocation Experiments: An Approach Based on Aggregate Data, *Journal of Marketing Research* 1983, **20**, 350-67.

MARK TL, SWAIT J. Using Stated Preference and Revealed Preference Modeling to Evaluate Prescribing Decisions, *Health Economics*, 2004, **13**, 563-73.

MASON C, QUIGLEY JM. Comparing the Performance of Discrete Choice and Hedonic Models, in: *Spatial Choices and Processes. Studies in Regional Science and Urban Economics, Vol. 21*, M. M. Fischer, P. Nijkamp und Y. Y. Papageorgiou (eds.), Amsterdam, Oxford, and Tokyo: North-Holland, 1983, 219-46.

MCFADDEN D. Conditional Logit Analysis of Qualitative Choice Behavior, in: *Frontiers in Econometrics*, Zarembka P (ed.), New York: Academic Press, 105-42.

MOUNT T, WENG W, SCHULZE W, *et al. Automobile Safety and the Value of Statistical Life in the Family: Valuing Reduced Risk for Children, Adults and the Elderly*, Working Paper, United States Environmental Protection Agency, 2000.

NOCERA S, BONATO D, TELSER H. The Contingence of Contingent Valuation – What Are People Willing to Pay against Alzheimer's Disease? *International Journal of Health Care Finance and Economics* 2002, **2**, 219-40.

RYAN M. A Comparison of Stated Preference Methods for Estimating Monetary Values, Health Economics 2004, **13**, 291-96.

RYAN M, GERARD K. Using Discrete Choice Experiments to Value Health Care Programmes: Current Practice and Future Reflections, *Applied Health Economics and Health Policy* 2003, 2(1), 55-64.

SCHOENBERG R, ARMINGER G. *Linear Covariance Structures Version 2.0, User Guide, RJS Software,* Kensigton, Maryland 1989.

SHEPARD DS, ZECKHAUSER RJ. Life-Cycle Consumption and Willingness to Pay for Increased Survival, in: *The Value of Life and Safety*, M. W. Jones-Lee (ed.), Amsterdam, New York, Oxford: North-Holland Publishing Company, 1982, 95-141.

SOMMER H, SEETHALER R, CHANEL O, *et al. Health Costs Due to Road Traffic-Related Air Pollution: An Impact Assessment Project of Austria, France and Switzerland*, Bern: Eidgenössisches Departement für Umwelt, Verkehr, Energie und Kommunikation, prepared for the WHO Ministerial Conference on Environment and Health, London June 1999.

TELSER H, ZWEIFEL P. Measuring Willingness-to-Pay for Risk Reduction: An Application of Conjoint Analysis, *Health Economics* 2002, **11**, 129-39.

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY. *The Benefits and Costs of the Clean Air Act 1990 to 2010*, EPA Report to the Congress, Report Nr. EPA-410-R-99-001, 1999.

VISCUSI WK. *Fatal Tradeoffs: Public and Private Responsibilities for Risk*, New York: Oxford University Press, 1992.

ZWEIFEL P. On the Use of Willingness-to-Pay Studies in Health, Schweizerische Zeitschrift fur Volkswirtschaft und Statistik / Swiss Journal of Economics and Statistics 2001, **137**, 11-25.

ZWEIFEL P, BREYER F. *Health Economics*, New York, Oxford: Oxford University Press, 1997.

Working Papers of the Socioeconomic Institute at the University of Zurich

The Working Papers of the Socioeconomic Institute can be downloaded from http://www.soi.unizh.ch/research/wp/index2.html

0505	Consumer Resistance Against Regulation: The Case of Health Care Peter Zweifel, Harry Telser, and Stephan Vaterlaus, February 2005, 23p.
0504	A Structural Model of Demand for Apprentices Samuel Mühlemann, Jürg Schweri, Rainer Winkelmann and Stefan C. Wolter, February 2005, 25p.
0503	What can happiness research tell us about altruism?Evidence from the German Socio-Economic Panel Johannes Schwarze and Rainer Winkelmann, February 2005, 26p.
0502	Spatial Effects in Willingness-to-Pay: The Case of Nuclear Risks Peter Zweifel, Yves Schneider and Christian Wyss, January 2005, 37p.
0501	On the Role of Access Charges Under Network Competition Stefan Buehler and Armin Schmutzler, January 2005, 30p.
0416	Social Sanctions in Interethnic Relations: The Benefit of Punishing your Friends Christian Stoff, Dezember 2004, 51p.
0415	Single Motherhood and (Un)equal Educational Opportunities: Evidence from Germany Philippe Mahler and Rainer Winkelmann, November 2004, 23p.
0414	Are There Waves in Merger Activity After All? Dennis Gärtner and Daniel Halbheer, September 2004, 39p.
0413	Endogenizing Private Information: Incentive Contracts under Learning By Doing Dennis Gärtner, September 2004, 32p.
0412	Validity and Reliability of Willingness-to-pay Estimates: Evidence from Two Overlapping Discrete-Choice Experiments Harry Telser, Karolin Becker and Peter Zweifel. September 2004, 25p.
0411	Willingness-to-pay Against Dementia: Effects of Altruism Between Patients and Their Spouse Caregivers Markus König und Peter Zweifel, September 2004, 22p.
0410	Age and Choice in Health Insurance: Evidence from Switzerland Karolin Becker and Peter Zweifel, August 2004, 30p.
0409	Vertical Integration and Downstream Investment in Oligopoly Stefan Buehler and Armin Schmutzler, July 2004, 30p.
0408	Mergers under Asymmetric Information – Is there a Leomons Problem? Thomas Borek, Stefan Buehler and Armin Schmutzler, July 2004, 38p.
0407	Income and Happiness: New Results from Generalized Threshold and Sequential Models Stefan Boes and Rainer Winkelmann, June 2004, 30p.
0406	Optimal Insurance Contracts without the Non-Negativity Constraint on Indemnities Revisited Michael Breuer, April 2004, 17p.
0405	Competition and Exit: Evidence from Switzerland Stefan Buehler, Christian Kaiser and Franz Jaeger, March 2004, 28p.
0404	Empirical Likelihood in Count Data Models: The Case of Endogenous Regressors Stefan Boes, March 2004, 22p.
0403	Globalization and General Worker Training Hans Gersbach and Armin Schmutzler, February 2004, 37p.

0402 Restructuring Network Industries: Dealing with Price-Quality Tradeoffs
Stefan Bühler, Dennis Gärtner and Daniel Halbheer, January 2004, 18p.

0401 Deductible or Co-Insurance: Which is the Better Insurance Contract under Adverse Selection?
Michael Breuer, January 2004, 18p.

0314 How Did the German Health Care Reform of 1997 Change the Distribution of the Demand for Health Services?
Rainer Winkelmann, December 2003, 20p.

0313 Validity of Discrete-Choice Experiments – Evidence for Health Risk Reduction
Harry Telser and Peter Zweifel, October 2003, 22p.

0312 Parental Separation and Well-Being of Youths
Rainer Winkelmann, October 2003, 20p.

0311 Re-evaluating an Evaluation Study: The Case of the German Health Care Reform of 1997
Rainer Winkelmann, October 2003, 23p.

0310 Downstream Investment in Oligopoly
Stefan Buehler and Armin Schmutzler, September 2003, 33p.

0309 Earning Differentials between German and French Speakers in Switzerland
Alejandra Cattaneo and Rainer Winkelmann, September 2003, 27p.

0308 Training Intensity and First Labor Market Outcomes of Apprenticeship Graduates
Rob Euwals and Rainer Winkelmann, September 2003, 25p.

0307 Co-payments for prescription drugs and the demand for doctor visits – Evidence from a natural experiment
Rainer Winkelmann, September 2003, 22p.

0306 Who Integrates?
Stefan Buehler and Armin Schmutzler, August 2003, 29p.

0305 Strategic Outsourcing Revisited
Stefan Buehler and Justus Haucap, July 2003, 22p.

0304 What does it take to sell Environmental Policy? An Empirical Analysis for Switzerland
Daniel Halbheer, Sarah Niggli and Armin Schmutzler, 2003, 30p.

0303 Mobile Number Portability
Stefan Buehler and Justus Haucap, 2003, 12p.

0302 Multiple Losses, Ex-Ante Moral Hazard, and the Non-Optimality of the Standard Insurance Contract
Michael Breuer, 2003, 18p.

0301 Lobbying against Environmental Regulation vs. Lobbying for Loopholes
Andreas Polk and Armin Schmutzler, 2003, 37p.

0214 A Product Market Theory of Worker Training
Hans Gersbach and Armin Schmutzler, 2002, 34p.

0213 Weddings with Uncertain Prospects – Mergers under Asymmetric Information
Thomas Borek, Stefan Buehler and Armin Schmutzler, 2002, 35p.

0212 Estimating Vertical Foreclosure in U.S. Gasoline Supply
Zava Aydemir and Stefan Buehler, 2002, 42p.

0211 How much Internalization of Nuclear RiskThrough Liability Insurance?
Yves Schneider and Peter Zweifel, 2002, 18p.

0210 Health Care Reform and the Number of Doctor Visits ? An Econometric Analysis
Rainer Winkelmann, 2002, 32p.

0209 Infrastructure Quality in Deregulated Industries: Is there an Underinvestment Problem?
Stefan Buehler, Armin Schmutzler and Men-Andri Benz, 2002, 24p.