AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2013WR014905

Key Point:
- Benefit transfer reliability for water quality improvement estimates

Supporting Information:
- Appendices
- Supporting figures A1–A5

Correspondence to:
S. Ferrini,
s.ferrini@uea.ac.uk

Citation:
Ferrini, S., M. Schaafsma, and I. Bateman (2014), Revealed and stated preference valuation and transfer: A within-sample comparison of water quality improvement values, *Water Resour. Res.*, *50*, doi:10.1002/2013WR014905.

Received 16 OCT 2013
Accepted 8 MAY 2014
Accepted article online 12 MAY 2014

Revealed and stated preference valuation and transfer: A within-sample comparison of water quality improvement values

Silvia Ferrini[1,2], Marije Schaafsma[1], and Ian Bateman[1]

[1]Centre for Social and Economic Research on the Global Environment, School of Environmental Sciences, University of East Anglia, Norwich, UK, [2]Department of Political and International Science, University of Siena, Siena, Italy

Abstract Benefit transfer (BT) methods are becoming increasingly important for environmental policy, but the empirical findings regarding transfer validity are mixed. A novel valuation survey was designed to obtain both stated preference (SP) and revealed preference (RP) data concerning river water quality values from a large sample of households. Both dichotomous choice and payment card contingent valuation (CV) and travel cost (TC) data were collected. Resulting valuations were directly compared and used for BT analyses using both unit value and function transfer approaches. WTP estimates are found to pass the convergence validity test. BT results show that the CV data produce lower transfer errors, below 20% for both unit value and function transfer, than TC data especially when using function transfer. Further, comparison of WTP estimates suggests that in all cases, differences between methods are larger than differences between study areas. Results show that when multiple studies are available, using welfare estimates from the same area but based on a different method consistently results in larger errors than transfers across space keeping the method constant.

1. Introduction

The European Union Water Framework Directive (WFD) requires all EU member states to achieve "good ecological status" in their water bodies by 2015. This ambitious target will have effects on environmental quality and human wellbeing. However, the cost of achieving these targets is expected to be substantial at £2.4 billion per year over a 43 year period just in England and Wales [*Metcalfe et al*., 2012]. The WFD allows exemptions from implementation of improvements in case of "disproportionate costs" [*RPA*, 2003]. This latter requirement mandates the use of methods for valuing water quality. EU States must provide reliable welfare measures for justifying that economic benefits are smaller than costs for exemptions to be granted.

The changes to the quality of open access waters (such as rivers and lakes) typically generate public goods benefits, many of which are nonmarketed. As such, analysts wishing to value these changes have to rely upon nonmarket valuation techniques. These can be broadly subdivided into two groups. Revealed preference (RP) methods, such as the travel cost (TC) approach, infer values from observed behavior, and thus measure use values [*Champ et al*., 2003]. Stated preference (SP) methods, such as the contingent valuation (CV) method, attempt to elicit values by asking direct valuation questions to respondents via surveys [*Bateman et al*., 2002]. SP approaches ask the individual to make choices in a hypothetical market between scenarios that can be formulated to identify both use and nonuse values [e.g., *Metcalfe et al*., 2012].

While both RP and SP methods have a long history of applications, recent years have seen an increase in the use of the Benefit Transfer (BT) approach [*Boyle et al*., 2010; *Johnston and Rosenberger*, 2010]. This method is promising when time and resource constraints bind cost-benefits analysis to incorporate nonmarket values. In this case, the benefit or cost assessment of nonmarket goods are "transferred" across space from a surveyed "study site" to a "policy site" [*Desvousges et al*., 1992]. The main advantage of applying BT methods is the cost saving on primary data collection, but this comes at the expense of the uncertainty about the accuracy of the transfer results.

The selection of primary estimates is crucially important for the reliability of the BT results. To reduce transfer errors and increase the level of acceptance of the results, primary studies used for BT can be evaluated

and selected based on a number of characteristics and quality criteria [*Boyle and Bergstrom*, 1992]. First, both the environmental good under valuation and the characteristics of the population whose welfare relies on these goods should be similar at the study and policy sites. Second, the valuation methods applied in the primary study, the reliability and validity of the results, and their acceptance among decision-makers are relevant selection criteria.

Compared with RP studies that rely on data of actual behavior, SP studies are regularly criticized among academics and policy-makers for their hypothetical nature. SP studies can capture nonuse values, but they rely on hypothetical statements that may fail to account for substitutes and budget limitations or be biased by interviewer-effects. Yet, as argued by *Azavedo et al*. [2003], Willingness To Pay (WTP) results of TC may be also biased, because the trip prices are not revealed by the respondent but determined by the analyst. Both methods hence have their limitations, but nonetheless one would expect a certain level of similarity in welfare estimates as well as model results.

In a BT exercise, the question that an analyst may be faced with is which primary study would be most reliable to use from a set of multiple potentially suitable valuation studies. For example, when TC results are preferred for policy-acceptance of a cost-benefit analysis yet the only available primary studies are a CV study from the policy area and TC studies from another study area. Would it then be better to use a TC study from another area, or are the differences between CV and TC both from the same area smaller?

Our paper aims to contribute to the discussion about BT reliability and validity using two valuation methods: the CV and TC methods. It provides an application of the BT method for valuing water quality improvements in the UK. Few recent applications of BT based on TC exist in the published academic literature, and there seem to be no papers using TC to estimate the welfare implications of WFD implementation. This paper is the first study to use TC for valuation of WFD benefits and to assess the transfer errors of these TC welfare estimates. We compare CV and TC results and provide an extensive test of BT reliability at a regional level. We investigate the validity of different BT procedures using and contrasting the CV and TC data as the primary source of information. Relying on more than 1700 observations collected face-to-face in the North of England in 2008, the paper aims at:

1. Testing the convergent validity of CV and TC results;
2. Testing the convergent validity of transferred values;
3. Comparing the transferability of CV and TC data;
4. Contrasting transfers across space and methods.

The paper is organized as follows. Section 2 gives a summary of the main findings of BT studies for water quality changes in Europe. Section 3 briefly sets out the case study, and describes the modeling approaches that will be used to analyze the data and the main hypotheses that will be tested. Section 4 first gives a summary of the main findings, followed by a comparison of the estimated WTP values and their confidence intervals. In section 5, different BT procedures are tested and the main results are discussed. Conclusions and implications for future BT studies are provided in the last section.

2. Benefit Transfer of WTP Values for Water Quality Changes

BT and its reliability depend, above all, on the availability of robust primary valuation studies. The economic valuation literature on water quality changes has a long history, especially in the United States [e.g., *Boyle et al*., 1993; *Kirchhoff et al*., 1997]. In Europe, this literature contains multiple papers related to the European WFD that came into force in the year 2000 [e.g., *Meyerhoff et al*., 2014; *Schaafsma et al*., 2012; *Kataria et al*., 2012; *Hanley et al*., 2006a, 2006b]. The majority of these applications use SP methods, thereby capturing the nonuse value related to water quality improvements under the WFD [*Brouwer*, 2008]. As this body of literature grows, so do the possibilities for BT.

The validity of BT results can be tested when primary studies for two sites are available so that the transferred WTP estimates can be compared to actual WTP estimates using commonly applied statistical equivalence tests [e.g., *Kristofersson and Navrud*, 2005; *Brouwer*, 2000]. Tests of convergent validity provide a useful test of primary data, and can be used as a selection criterion for studies informing BT.

The convergent validity of primary WTP estimates can be assessed by comparing WTP estimates to the results of other studies using the same or different elicitation formats and methods [*Rolfe and Dyack*, 2010; *Loomis et al*., 1991]. From a theoretical point of view, WTP that have been elicited with different CV question formats should be similar, but empirical studies have found a systematic difference between Dichotomous Choice (DC), Payment Card (PC) and Open-ended formats, where DC studies typically produce higher WTP estimates [*Cameron et al*., 2002, *Ready et al*., 1996; *Brown et al*., 1996]. The source of this bias may, among others, be "yea-saying" behavior in CVDC responses that have been elicited in face-to-face interviews. CVDC formats may be more prone to this type of bias than CVPC surveys [*Ready et al*., 1996; *Kealy and Turner*, 1993]. On the other hand, CVDC is considered to be potentially incentive compatible and recommended by the NOAA panel.

CV and TC estimates are also expected to be comparable and satisfy the convergent validity test when used to assess the same type of values for the same good, despite the differences between the methods. TC values are based on RP data of actual behavior and do not reflect nonuse values. Nonuse values can be assessed with SP methods using hypothetical markets. TC estimates are therefore expected to be lower than CV estimates for goods that have nonuse value. The empirical evidence is not conclusive on this matter. *Mbewaze and Bennett* [2012], *Loomis et al*. [1991], and *Smith et al*. [1986] compared TC and CV models, but did not find significant differences between marginal CVPC and TC-based WTP estimates. *Rolfe and Dyack* [2010] showed that TC generates higher estimates than CV and argued that a combination of factors is likely to drive systematic variation in WTP. Some meta-analyses show differences between valuation methods, with TC generally providing lower economic value estimates than SP studies [e.g., *Shrestha and Loomis*, 2003; *Brander et al*., 2006]. When significant differences between WTP values are found, it is impossible to decide which value is closest to the "true" WTP given the limitations inherent in both SP and RP methods [*Azavedo et al*., 2003].

The BT literature presents three main procedures to transfer estimates. The Unit value transfer (UVT) transfers the mean WTP estimates from the primary "study site" to the "policy site" assuming that the two populations are similar and the good under valuation is identical. While these assumptions are easily violated, many agencies are using the UVT approach for its simplicity. The Adjusted value transfer (AVT) takes into account significant differences between the study and policy site populations. The mean welfare measures can be adjusted by several factors, among others, income. The income adjustment has been found successful in different international applications and is less relevant for regional (within-country) BT where the two populations of interest are very similar. The Function transfer (FT) approach also takes into account possible differences between the study and policy sites and populations. The FT approach uses the model parameters obtained at the study site and applies these to secondary data available at the policy site.

The evidence in the empirical literature does not point to a single optimal approach to transfer economic values for water quality improvements. *Barton* [2002] tests the validity of BT at national level of CV-based values for water quality improvements in Costa Rica, and finds that UVT and AVT both provide acceptable (lower than 30%) transfer errors. In a BT exercise at national level using CV values for lake water quality improvements in Germany, *Muthke and Holm-Mueller* [2004] propose a *t* test that takes into account the transfer error that the decision-maker is willing to accept. Their results suggest that the AVT approach outperforms FT. In their case study at the national level, the BT error is generally lower than a chosen tolerated error of 50%, even though commonly applied statistical equivalence tests of BT hypothesis (*t* tests and likelihood ratio tests) are rejected. The study also tests the transferability of WTP estimates for water quality between Germany and Norway, but none of the three BT approaches results in acceptable transfer errors (here, $<$60%). *Bateman et al*. [2011] present a BT study directly related to the WFD. Using the results of an international CV study specifically designed to estimate the potential welfare changes of the implementation of the WFD, *Bateman et al*. [ibid.] show that BT across countries is most reliable when using FT, adjusting values for purchasing power parity (PPP).

Few if any recent academic studies compare different BT methods using TC data, and there are no TC published studies on the public benefits of the WFD. Older work in the United States by *Loomis* [1992] shows that UVT of TC data produces higher errors than FT. *Loomis et al*. [1995] find similar results for their within-sample test of BT for TC data.

3. Case Study and Hypotheses

3.1. Survey

To compare the transferability of economic values for water quality changes based on RP and SP valuation methods, we developed a questionnaire that included a TC assessment and a CV question, with either a DC or a PC format. Our study area is part of the Humber Catchment in the North of England (Figure 1). The three main rivers in this area, the rivers Wharfe (north), Aire (middle), and Calder (south), vary greatly in their water quality. The sampling area is expected to be relatively homogeneous in terms of cultural and geographic characteristics but different in socio-economic and environmental characteristics. These differences allowed constructing a number of tests of BT reliability.

In 2008, a large group of residents in the area were interviewed in person by enumerators using a computerized questionnaire. Interviewers had been carefully trained to use the system and follow the survey guidelines defined for the study. The sampling strategy was built on an efficient sample aimed at maximizing the statistical fit of the model to be estimated. This sampling distribution resulted in 26 sampling locations scattered across the sampling area, and therefore, maximized the spatial variation. Appendix SI provides more details on the sample and questionnaire design.

At the beginning of the questionnaire, a touch screen map of the study area (covering approximately 80 km^2) and the three rivers as well as all other rivers was presented. On this map, respondents were asked to identify the location of their home and the river sites that they had visited in the last 12 months and the number of visits to these sites. The sites were later matched to a real world recreational site using Geographical Information Systems (ArcGIS 9.2, ESRI) software. The distance by road, and travel time by car, from each respondent's home to all of the 531 recreational sites was calculated in the GIS. The 531 sites were aggregated to 84 river stretches of 5 km each for the purpose of statistical estimation of the TC data and comparison to the CV scenario.

In the CV-part of the survey, respondents were given a description of the current water quality and a map showing how the quality varied across the rivers (see Figure 1, left). The current water quality was calculated from Environment Agency long-term water quality monitoring data and categorized as follows [*Hime et al.*, 2009]:

1. Good (blue-pristine level),
2. Medium (green-good level),
3. Poor (respectively, yellow or red).

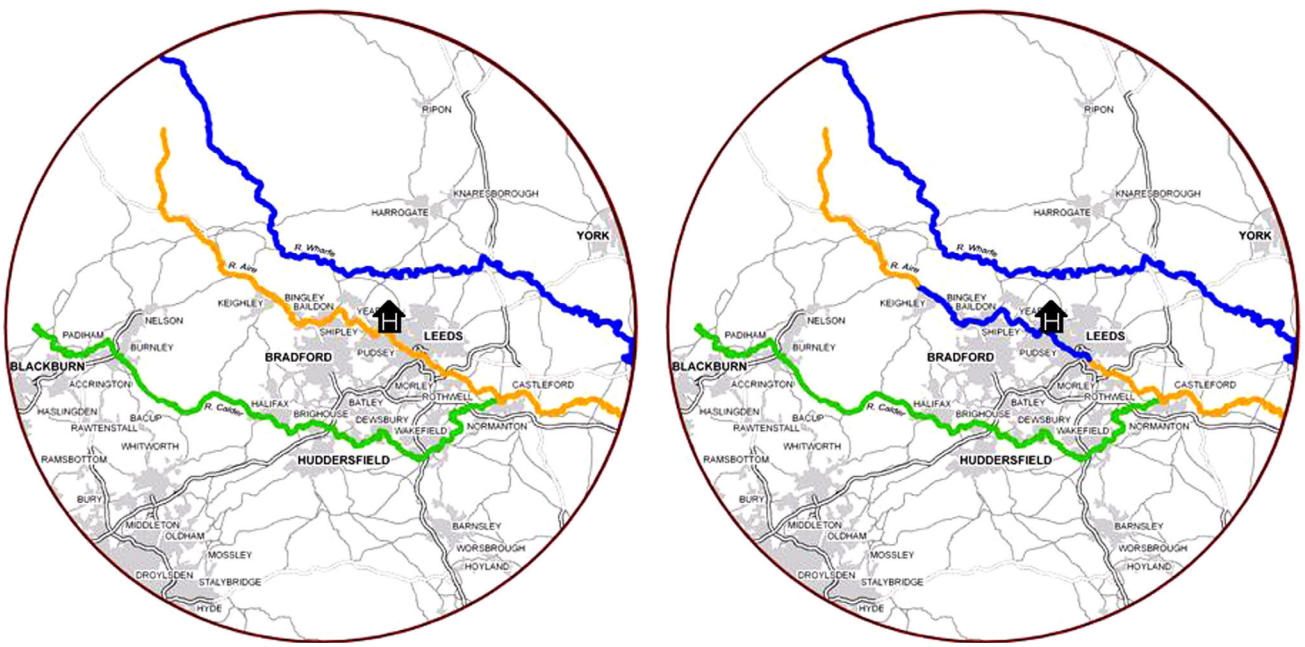

Figure 1. Maps of the study area: (left) the current situation and (right) the proposed water quality change. The map's diameter is roughly 70 km.

After in-depth discussions with ecologists and biologists, this scale was translated into a so-called "water quality ladder" that associated each level with different ecological and recreational amenities to give a meaningful explanation of the quality levels to respondents [*Hime et al*., 2009]. A graphic designer developed pictures for each quality level (Figure SA1, Appendix SI), reflecting the water clarity and the condition of plants and animals, together with pictograms indicating the possibilities for recreational use of the water body (bathing and fishing) and a textual description.

Next, a scenario of water quality improvements was offered, in which a stretch of the river Aire of roughly 20 km would be improved from "poor" to "good" quality (see Figure 1, right).

The CV evaluation question was posed by asking whether the survey participant would accept an increase in annual water bill in exchange for the proposed improvement in water quality. The choice of the water bill as payment mechanism was tested in the pilot studies and followed previous studies on water quality improvements in public areas [e.g., *Bateman et al*., 2011; *Willis et al*., 2005]. In the UK, the water bill covers water management costs as well as environmental and water quality maintenance.

Based on the water quality ladder, scenarios and payment vehicle, the CV-WTP answers could reflect use and nonuse values. Two different elicitation formats were used: respondents received either PC or DC questions (see Appendix SI for details). The payment card contained 60 different values ranging from £0 to £905, and a "do not know" option. The payment card was used to support respondents in answering an open-ended question, similar to the application in, for example, *Ready et al*. [1996]. Respondents could express any amount that reflected their maximum WTP. The resulting WTP data are hence continuous. The bid vector of the DC version ranged from £10 to £140 using seven points. The WTP questions were followed by a number of questions to discriminate between protest and valid zero bids. Socio-economic data were collected at the end of the questionnaire.

3.2. Modeling Approach

Standard methods for regression analyses for both CV and TC data were employed. The theory underlying the modeling procedures is well outlined in *Haab and McConnell* [2002]. The estimation of WTP models that are suitable for BT purposes reduces options for selection of explanatory variables as secondary data should be available for these variables at the policy site. Moreover, *Bateman et al*. [2011] find that WTP models that include theoretically derived variables produce lower BT errors than statistically best fit models. The authors argue that the inclusion of nontheoretically derived ad hoc variables tends to increase statistical fit in survey areas by picking up locally important factors. However, large errors can arise if those factors are not representative for the policy areas. Economic theory suggest only a limited set of factors that should apply universally, including respondent income, the costs of using a resource (proxied by distance) and the cost of using substitutes. In this paper, only explanatory variables that have a sound theoretical basis and for which secondary data from the study area are available were selected.

For the CVPC data, the dependent variable is continuous, and therefore, a Tobit model is used. In this model, the WTP values are considered as the latent variable in the modeling function [see *Santagata and Signorello*, 2000; *Kirchhoff et al*., 1997]. The WTP-function is specified as:

$$WTP_i = \beta_0 + \beta_1 Inc_i + \beta_2 Dist_i + \beta_3 Dist\_Sub_i \quad (1)$$

where the β parameters are to be estimated. Income (*Inc*) is expected to be positively correlated to WTP reflecting a higher ability to pay. The distance between the respondent and the stretch of river under valuation (*Dist*) is expected to have a negative effect, because higher travel distances are associated with higher costs of visiting and hence lower utility. The distance to substitutes (*Dist_Sub*) reflects the costs of visiting other rivers and the relative scarcity of the offered good, and should therefore have a positive impact in the WTP model according to economic theory. The CVPC WTP estimates are obtained using the model estimates, following standard procedures for the Tobit model described in *Kirchhoff et al*. [1997].

For the CVDC data, we follow *Hanemann* [1989] and estimate a function similar to equation (1). We estimate a logit model in which a "Yes" response is assumed to reflect a WTP equal or higher than the bid amount, and a "No" response otherwise (see equation (2)):

$$P(yes) = \beta_0 + \beta_1 Bid_i + \beta_2 Dist_i + \beta_3 Dist\_Sub_i. \quad (2)$$

In the DC modeling framework, the bid parameter β_1 reflects the marginal utility of income changes. The probability that a respondent will say "Yes" is expected to decrease with the bid level and the distance to the river stretch under improvement, and to increase as the distance to substitutes increases.

In the TC analysis, we apply a standard Random Utility Model approach and adopt a repeated Nested Logit Model (NL) as in *Morey et al.* [1993]. In the TC model, the utility of visiting a river site depends on the costs of visiting river sites, river water quality at the site, and the availability of other nonriver alternatives. The choice is modeled whether and where to go for water recreation over a period of 1 year, assuming that each day represents a choice occasion ($t = 1,\ldots, 365$). As suggested in the literature [e.g., *Hynes et al.*, 2009; *Bockstael and McConnell*, 2007; *Smith et al.*, 1983], travel cost are defined as the out-of pocket cost (using £0.25/km as the fuel cost) plus the opportunity cost of time calculated as a percentage of respondent's wage. Recent studies suggest that different proportions of wage may better reflect the value of travel time for recreation purposes [*Fezzi et al.* 2014; *Wolff*, 2014] but in this study the standard 1/3 of the respondent's wage is used. Under the assumption of individual utility maximization, respondent i at choice occasion t will choose the location j which provides the highest utility in that period given J possible recreation options. Set J ($j = 0,\ldots,88$) includes the 84 river stretches of ~5 km along the three rivers shown in Figure 1, the option not to recreate ($j = 0$), and trips to other water recreation sites for which annual trip numbers were collected, such as canals, lakes, and rivers not in the sampling area ($j = 85,\ldots,88$). In a RUM framework, individual utility is a function of river sites water quality, alternative specific constants, and travel costs:

$$U_{ijt} = \eta_j Asc_j + \beta_m X_{mj} + \beta_p X_{pj} + \gamma Tc_{ij} + \varepsilon_{ijt}, \quad (3)$$

where η, γ, and β are parameters to be determined, X_{mj} and X_{pj} are dummy variables which take the value 1 if water is medium (m) or poor quality (p), respectively. The *Asc* is a constant specific for option $j < 85$. *Tc* is the travel cost and its parameter (γ) represents the marginal utility of income. The error term ε_{ijt} reflects the utility of each respondent that remains unobservable to the researcher and is assumed to vary across periods t, options j, and individuals i. The error terms are assumed to be drawn from the generalized extreme value distribution similar to other studies of recreational choices which have used a two-branch NL model [e.g., *Lew and Larson*, 2008; *Needelman and Kealy*, 1995]. One branch includes all river sites along the three main rivers in our study area (*River*), whereas other recreation options and the option to stay home ($j = 0$, 85, . . ., 88) are included in the other branch (*No River*). At the lowest level of the NL model, the probability of choosing alternative k (one of the 84 river stretches), conditional on the choice of visiting a river site in our study area, is:

$$P[k|River] = \frac{e^{\boldsymbol{\beta}'\mathbf{x}_k + \gamma Tc_{ik}}}{\sum_j e^{\boldsymbol{\beta}'\mathbf{x}_j + \gamma Tc_{ij}}}. \quad (4)$$

The probability of choosing other recreation sites or stay home, conditional on the choice of not visiting river sites in the area, is:

$$P[h|No\ \ River] = \frac{e^{\eta_h Asc_h}}{\sum_w e^{\eta_w Asc_w}}. \quad (5)$$

where Asc_h represents the alternative specific constant associated with the option h. The frequency of visiting other water recreation options is used as baseline value, to which the other Asc parameters can be compared. For example, a positive estimate for the alternative specific constant associated with the option to stay home implies that on average respondents derive higher utility from staying home than from visiting water sites.

At the highest level of the NL model, the probability of choosing to visit river sites, as opposed to staying at home or going to other recreation sites, is driven by the expected maximum utility obtained from the lower level options of each branch:

$$P[River] = \frac{e^{IV_{River}}}{e^{IV_{River}} + e^{IV_{NoRiver}}}. \quad (6)$$

This probability is driven by the term IV (Inclusive Value), which is the logarithm of the sum of the expected maximum utility that the respondent will get from the lower level options of branches. The utility of lower level options are described in equations (4) and (5). The IV summarizes the utility of all options in one branch as the maximum expected benefit that the respondent will get by choosing either the *River* or the *No River* branch. The estimated TC model results are used to obtain welfare estimates for the same water quality improvement as proposed in the CV scenario (Figure 1, right), following standard procedures described in *Bockstael and McConnell* [2007].

The TC model is based on choices across a large number of sites, and captures substitution effects between river sites when quality and distance change. Other site characteristics have been excluded to improve the comparability of TC CV estimates. Moreover, secondary data are usually unavailable for the characteristics of policy site or require too much GIS information.

3.3. Split-Sample Approach

The transfer reliability tests are based on a novel variant of classic two sample testing approaches. Typically assessments of the validity of BT techniques gather data from at least two sites by conducting valuation exercises at both locations. One site is then designated as the "study" site and results used to transfer a value to the other "policy" site. Comparison of the transfer valuation with that actually conducted at the policy site forms the basis of the BT testing protocol. The direction of the test is then reversed and a further assessment of transferability conducted. Ideally transfers should take into account any variation in the socio-economic characteristics and substitute availability characterizing the two samples. However, if there is also variability in the good across the sites (e.g., the change in water quality is not identical) then a clear problem of confounding can arise, a problem which can be exacerbated by temporal variation between survey dates [*Brouwer and Bateman*, 2005]. Our study offers a novel variant on BT testing by applying the same survey instrument to the same study site facing an identical change in provision, but we assess the economic value of this change across two samples drawn from populations with differing socio-economic characteristics and substitute availability. Such a design controls for many of the potential confounding factors facing BT tests allowing us to focus upon remaining differences between the two populations.

In this case study, two subsamples of comparable size are created. Respondents living north of the river Aire are assigned to the North sample. This subsample includes 683 respondents, which is similar to the size of other primary studies used in BT applications [*Johnston et al*., 2006]. The remaining 1076 observations belong to the South sample. The difference in average river site quality between the North and South sampling areas allows examining whether objective differences in environmental quality are reflected in valuations. River quality in the North is good (blue) and 98% of river sites along the river Wharfe are already of pristine quality. In the South, 35% of river sites along the river Calder are of medium or poor quality.

In testing the convergent validity of the BT, the WTP estimates from the study site are transferred to the policy site and compared with the actual WTP estimates. Besides a geographical split, the sample is also split by elicitation format, creating one subsample for the TC method, one for the CVPC and one for the CVDC format. The CV questions elicited WTP per household and are divided by the number of adults in the household to allow for a comparison between CV and TC. The convergent validity of the primary data sets across methods is tested.

3.4. Tests of WTP Equivalence and BT Errors

3.4.1. Convergent Validity of Primary Data Sets

As outlined above, four related research objectives will be tested statistically. The first objective is to test the convergent validity of the primary SP and RP results. This comparison of values for the same location obtained through different methods provides information about the reliability of the primary data sets for use in BT. The hypothesis tested is:

$H_0 1 : WTP_{DC} = WTP_{PC} = WTP_{TC}$

H1 is tested for the full sample as well as the North and South subsamples. As a criterion for convergent validity, the overlap of the confidence intervals (CIs) of the welfare estimates is used. The three methods

should theoretically result in similar WTP estimates when the same good and values are assessed. In this paper, the CV estimates are expected to be greater or equal to TC estimates, because the former may also capture nonuse values, and empirical findings suggest that DC values may be higher than PC values [e.g., *Cameron et al*., 2002].

3.4.2. Benefit Transfer Tests

The split-sample approach allows testing the BT validity across space and valuation methods. Transfer errors are computed by comparing transferred and observed WTP: % BT error = (WTP (study site) − WTP (policy site))/WTP (policy site) * 100%. Following BT standards [e.g., *Kirchhoff et al*., 1997], the WTP estimates are obtained using mean parameter estimates ($\hat{\boldsymbol{b}}$) and the sample distribution for the explanatory variables (X_i). For the different transferability comparisons, the North WTP values and models are estimated and then transferred to the South as the policy site, and then vice versa, using both FT and UVT. The equality of the original and transferred WTP estimates is tested using a standard parametric two side *t* test [*Brouwer*, 2000]:

For UVT: $H_0 2a$: $\hat{WTP}_N = \hat{WTP}_S$

For FT: $H_0 2b$: $\hat{WTP}_N = \hat{WTP}_{\beta_S X_N}$ and $\hat{WTP}_S = \hat{WTP}_{\beta_N X_S}$.

The nonparametric Mann-Whitney statistic, which unlike the *t* test is robust against the nonnormality assumption, is also used. In addition, Likelihood Ratio tests are performed on the equivalence of the beta parameters of the TC and CV models for the North and the South:

For FT: $H_0 3$: $\hat{\beta}_N = \hat{\beta}_S$

Finally, the tolerance test by *Kristofersson and Navrud* [2005] is applied using a 20% tolerance error as an acceptable level of differences in WTP:

For UVT and FT: $H_0 4$: $\hat{WTP}_N - \hat{WTP}_S < 20\%$.

These statistical tests are performed and the resulting transfer errors are compared to meet the second objective of this paper. The prior expectation is that for both TC and CV data, FT will result in lower BT errors than UVT. To respond to the third objective, differences in transfer errors between methods (TC or CV) are tested to assess which valuation data set results in the lowest transfer error across sites, i.e., between the North and South.

Finally, these results are combined to assess whether transfer errors are smaller across space or valuation methods, i.e., for values based on different methods obtained in the same sampling area, or based on similar methods using data from different sampling areas. The valuation literature may be expanding and the role of economics in decision-making rising, but there are still considerable gaps in the availability of primary valuation data sets. Hence, although users of BT results may have an a priori preference for a particular method, they may have to trade-off acceptability (preferred method) against reliability (lowest BT error) in case the preferred method has only been applied in different study areas and results in considerably larger transfer errors.

4. Results

4.1. Descriptive Statistics

The final data set contains CV and TC data from 1759 respondents. The main descriptive statistics of the sample are presented in Table 1. 61% of the CV respondents were identified as zero-bidders and 10% as protesters. Outliers were defined as respondents who choose an amount of £150 or more from the payment card ($n = 3$), and who lived more than 60 km from the site ($n = 3$). Protesters and outliers were removed from the data set.

Despite the proximity of the two subsamples, there are statistically significant differences between the North and South sample in the main socio-economic characteristics such as age, income, and household size (*t* test comparisons results in Table 1). The full sample hence reflects conditions that are typically found in BT exercises.

The significant difference in mean and median distance to the river under valuation shows that the spatial distribution of the respondents differs between the subsamples. Respondents in the North subsample live closer to the improved river stretch but also closer to other (good quality) substitutes.

Table 1. Main Statistics of Socio-economic Variables

Variable	Pooled Mean (St. Dev.)	South Mean (St. Dev.)	North Mean (St. Dev.)	*t* Test North = South (95%)
Gender (male=0; female=1)	0.43 (0.50)	0.43 (0.50)	0.44 (0.50)	Accept
Age	51 (18)	48 (18)	54 (19)	Reject
Income (gross in GBP/household/year)	21,545 (11,541)	20,118 (10,483)	23,794 (12,724)	Reject
Household size	2.6 (1.4)	2.7 (1.4)	2.5 (1.3)	Accept
Urban residence (1=urban; 0=rural)	0.73 (0.45)	0.76 (0.43)	0.68 (0.47)	Reject
Distance to river in scenario: mean (km)	13.7 (10.5)	15.8 (10.6)	10.2 (9.5)	Reject
Distance to river in scenario: median (km)[a]	12.5	14.3	8.0	
Distance to substitute: mean (km)	6.2 (4.8)	6.8 (4.6)	5.2 (4.9)	Reject
Sample size	1759	1076	683	

[a]Standard deviation and *t* tests are not available for position measures like medians.

4.2. CV Model Results and WTP Estimates

In the final sample, 48% of the respondents received the CVPC format and the remaining 52% the CVDC question. For sake of brevity, the results of the regression analyses of the PC and CV data can be found in Appendix SII, including the Tobit models for the CVPC data and three logit models for the CVDC data (Table SA2.1) of the full sample as well as the North and South subsamples. Income is highly significant in the CVPC models, which is an important indicator of the validity of the results. In the CVDC version, the bid parameter is significant and negative which implies that the number of "Yes" responses decreases as bid levels increase.

As expected, distance to the river stretch has a significant negative effect in the models for the full sample and the South subsample, but is not significant at 5% in the North subsample. A possible explanation is that the average distance to the improved river stretch in the North is relatively low and differences in travel costs to the site are therefore not the main determinant of preferences and WTP. Contrary to expectations, the parameter of the distance to the substitute is not significant in any of the models.

The results of the TC model show that all variables are statistically significant (see Table SA2.2 in Appendix SII). The travel cost parameter, reflecting the marginal utility of income, is significant and negative as expected. This implies that respondents prefer closer river stretches with a lower travel cost. The water quality parameters, compared to the baseline level (good water quality), are both negative and significant. Both in the full and North samples, the medium quality parameter is smaller than that of the poor quality, which implies that, as expected, respondents are less likely to visit poor quality sites. For valuation purposes, this result is interpreted to mean that people attach higher utility to improving poor quality river sites than medium quality sites. However, in the South, the opposite is found, suggesting that people care more about improving medium quality sites than poor quality sites.

4.3. Convergent Validity of Primary Studies Estimates

The convergence validity test of primary studies aims to evaluate the quality of the results and can be used to inform study selection for BT purposes. As a test of the convergence validity of the WTP estimates, we compare the WTP results of the different methods and elicitation formats. The significant and positive WTP estimates show that people care about water quality improvements. Mean annual WTP for an improvement from poor to good quality ranges from a minimum of £10 to a maximum of £27 (full sample results). Mean WTP estimates for TC are lower than the CVDC estimates but higher than CVPC estimates.

Table 2. Individual WTP Values and Confidence Intervals (in £(2008)/yr)

	Full Sample	North	South
$\hat{WTP}_{CVPC}$			
Mean	10.02	10.96	9.44
(CI 95%)[a]	(3.47, 19.33)	(5.43, 19.90)	(2.47, 20.52)
$\hat{WTP}_{CVDC}$			
Mean	27.19	36.61	38.62
(CI 95%)[a]	(9.98, 59.58)	(15.83, 63.45)	(18.45, 58.93)
$\hat{WTP}_{TC}$			
Mean	20.16	18.26	22.89
(CI 95%)[a]	(0.00, 91.54)	(0.00, 81.09)	(0.01, 88.08)

[a]Confidence intervals are drawn from empirical distributions of WTP. The explanatory variables in the procedure are based on sample values.

The results show that the CIs of the CVDC, CVPC, and TC WTP estimates overlap. This finding suggests that the WTP results pass the convergent validity test and $H_0 1$ cannot

Table 3. WTP Values for Transfer Tests

Function Transfer (FT)[a]		North	South
$WTP^{PC}_{\beta_S X_{IN}}, WTP^{PC}_{\beta_N X_{IS}}$	Mean	12.89	8.97
	St. dev.	5.46	4.02
$WTP^{DC}_{\beta_S X_{IN}}, WTP^{DC}_{\beta_N X_{IS}}$	Mean	42.33	34.32
	St. dev.	11.21	14.55
$WTP^{TC}_{\beta_S X_{IN}}, WTP^{TC}_{\beta_N X_{IS}}$	Mean	28.87	9.95
	St. dev.	36.86	23.28

[a]The WTP estimate for the North is based on the betas of the WTP model estimated for the South sample, with sample values from the North for the explanatory variables. For the South, the WTP estimate is based on the betas of the WTP model estimated for the North sample, with sample values from the South for the explanatory variables.

be rejected. This result is partially confirmed by a convolutions test as suggested by *Poe et al.* [1994] which shows that the distribution of WTP values in the North subsamples are similar, whereas some differences are found in the South subsample. However, performing this test is usually not an option for BT practitioners, when only the study report, but not the underlying primary data, of the original studies are available.

To verify whether the difference between CVPC and CVDC is caused by the distributional assumptions of the Tobit and logit models rather than differences in stated WTP, a synthetic DC data set using the PC responses is constructed following the approach by *Cameron and Hupper* [1991]. Results show that synthetic mean WTP values are similar to actual CVPC WTP results and therefore differences between the CV elicitation formats in this study are argued to be due to differences in respondent behavior.

4.4. Convergent Validity of Transferred Values

The convergent validity test of BT values is based on the comparison of the original and transferred values: values used for tests of UVT are presented in Table 2, while Table 3 presents the WTP estimates based on FT.

Table 4 reports the equivalence tests results and calculated transfer errors. Results show that the errors are below 20% for all CV estimates, ranging from 5% to 18%, while for TC, the transfer errors are 20% or higher with a maximum of 58%.

According to *t* test results $(H_0 2)$, the null-hypothesis of WTP equality of UVT results are rejected for TC and CVPC, and only just accepted for CVDC ($p = 0.054$). For the transfer of CVPC results from the North to the South, the null-hypothesis cannot be rejected. Nonparametric Mann-Whitney tests give similar results for the FT and reject all UVT comparisons.

The LR test used to test model equality $(H_0 3)$, shows that the CVPC models of the North and South have statistically similar parameters, but the equality of beta parameters of the CVDC and TC models is rejected.

The results of the tolerance test, with an acceptable error margin set at 20% $(H_0 4)$, provide a minor increase in support of transferability compared with the *t* test. In this case, UVT results of CVPC values from the South to the North are also associated with a tolerated error, and so are the FT results of CVDC values from the North to the South. The tolerance level has to be increased to 25% for all CV transfers to result in acceptable errors, and to 40% for TC UVT and to 80% for TC FT.

Table 4. Transfer Tests and Errors

			H2		H3	H4	
	Policy Site	Study Site	*t* Test (5%)	Mann-W. Test	LR Test	Tolerance Test (20%)	Transfer Error
CVPC							
UVT	$W\hat{T}P_N$	$W\hat{T}P_S$	Reject	Reject		Nonreject	16%
	$W\hat{T}P_S$	$W\hat{T}P_N$	Reject	Reject		Reject	14%
FT	$W\hat{T}P_{\beta_S X_{IN}}$	$W\hat{T}P_N$	Reject	Reject	Nonreject	Reject	18%
	$W\hat{T}P_{\beta_S X_{IS}}$	$W\hat{T}P_S$	Nonreject	Nonreject	Nonreject	Nonreject	5%
CVDC							
UVT	$W\hat{T}P_N$	$W\hat{T}P_S$	Nonreject	Reject		Nonreject	6%
	$W\hat{T}P_S$	$W\hat{T}P_N$	Nonreject	Reject		Nonreject	5%
FT	$W\hat{T}P_{\beta_S X_{IN}}$	$W\hat{T}P_S$	Reject	Reject	Reject	Reject	16%
	$W\hat{T}P_{\beta_S X_{IS}}$	$W\hat{T}P_S$	Reject	Reject	Reject	Nonreject	11%
TC							
UVT	$W\hat{T}P_N$	$W\hat{T}P_S$	Reject	Reject		Reject	20%
	$W\hat{T}P_S$	$W\hat{T}P_N$	Reject	Reject		Reject	25%
FT	$W\hat{T}P_{\beta_S X_{IN}}$	$W\hat{T}P_N$	Reject	Reject	Reject	Reject	58%
	$W\hat{T}P_{\beta_S X_{IS}}$	$W\hat{T}P_S$	Reject	Reject	Reject	Reject	57%

Table 5. Equivalence *t* Test and Transfer Errors Across Space or Methods[a]

Policy Site		North			South		
Preferred Welfare Estimate		CVPC	CVDC	TC	CVPC	CVDC	TC
Study Site	Available Primary Studies						
North	CVPC		Reject (234%)	Reject (67%)	Reject (14%)		
	CVDC	Reject (70%)		Reject (50%)		Nonreject (5%)	
	TC	Reject (40%)	Reject (101%)				Reject (25%)
South	CVPC	Reject (16%)				Reject (309%)	Reject (142%)
	CVDC		Nonreject (6%)		Reject (76%)		Reject (41%)
	TC			Reject (20%)	Nonreject (59%)	Reject (69%)	

[a]Errors for situations with variation across space and methods are not relevant to the objective of our study and have not been estimated. The two tests (*t* tests and tolerance tests at 20%) lead to similar results, except for the transfer of CVPC values from the South to the North where the tolerance test would not reject the equality of WTP estimates.

The comparison of transfer errors of UVT and FT shows that errors of UVT are generally lower than that of FT, especially for the TC data. The better performance of UVT in this case is surprising considering the significant differences in socio-demographic variables (Table 1) and current river water quality levels between the North and South.

4.5. Contrasting BT Performance Across Methods and Space

The split-sample design provides the opportunity to contrast the transferability of WTP estimates across space (here: from North to South and vice versa) and methods (from CVPC to CVDC, etc.). Table 5 presents *t* tests and transfer error results of BT across space or methods using UVT. The results of the upper right and lower left blocks of Table 5 suggest that UVT across space (using the same method) results in transfer errors below 16% for CV results, and errors of 20–25% for TC.

The upper left and lower right blocks present the *t* test results and transfer errors of UVT estimates from alternative valuation methods applied in the same area. For example, a policy-maker in the North, who prefers CVPC data to value water quality changes and has access to a CVDC and a TC study from the North and a CVPC study from the South, would obtain lower transfer errors using the CVPC study from the South (16%) than with the original studies from the North (70% or 40%).

In this study, all transfers across space (holding the method constant) outperform transfers across methods (where the policy and study site are the same). In case a policy-maker prefers using a primary study performed in the same area, then a trade-off has to be made between the accuracy of the transfer estimate against the acceptability of the original data source.

5. Conclusions

This paper presents the findings from a survey conducted in the North of England on the nonmarket benefits of the implementation of the WFD. A large sample was collected using CV and TC approaches. This is the first paper to provide a full in-sample test of convergence and transfer performance of TC results and to compare this with CV results. Moreover, it is the first paper providing welfare estimates related to the WFD based on TC data.

The welfare estimates show that respondents attach a positive WTP to improving water quality changes to meet the objectives of the WFD in the river Aire in the UK with a minimum value of £10 per year per person. Convergence validity test results show that the confidence intervals of WTP estimates based on CVPC, CVDC, and TC overlap. However, their performance in our in-sample BT application is different.

Statistical equivalence tests of transferred values show that there are significant differences across the two geographically distinct subsamples in WTP models and value estimates. Out of the three elicitation formats tested, the CV data (PC and DC) produce better BT results than the TC data, with transfer errors lower than 20% for both UVT and FT. CVPC models of the North and South sample also have comparable beta parameters. The difference in transfer accuracy of UVT and FT of CV values is minor and not systematic.

For the TC data, larger transfer errors are found. Transferring the results of the TC data from the North to the South (or vice versa) produces the transfer errors of 20%–25% for UVT, but over 50% for FT and equivalence tests of UVT and FT are rejected in all cases.

Further comparison of WTP estimates suggests that in all cases, differences in WTP between valuation methods are larger than between study areas. The results of this study suggest that when multiple studies are available, using welfare estimates from the same area but based on a different method consistently generates larger transfer errors than transfers across space keeping the method constant. Although the literature provides earlier evidence of differences in WTP values between methods, this study is the first to assess if such differences are indeed larger than those resulting from transferring values across space. It would be helpful for BT purposes if primary studies put more effort into convergence validity tests by comparing welfare estimates both across space and across methods.

In some contexts, policy-makers may be inclined to put more trust in results from one valuation method than from others. There is an on-going discussion in the literature about the reliability and validity of RP and SP studies and each technique has its limitations. In the RP analysis presented in this paper, like in most other TC studies, the travel costs are not observed, but inferred from respondents' stated income and travel mode and route assumptions, and there may be additional measurement and recall errors of the visitation frequencies across the available sites over a period of 1 year. On the other hand, the CV results are based on hypothetical markets where respondents are asked the unfamiliar question to price river water quality changes and its associated benefits in terms of informal recreation and biodiversity.

The tolerance level applied in this paper for equivalence tests is low compared with previous studies [e.g., *Muthke and Holm-Mueller*, 2004]. The error that a policy-maker is willing to accept can depend on the stage in the decision-making process that the BT exercise has to inform. For policy making at strategic levels, much higher tolerance levels could be applied, and the results show that with a 25% tolerance level all CV transfers produce acceptable results, and with a 40% tolerance level TC UVT would be acceptable too. Yet, toward the implementation phase of projects, especially for governance at regional and local scales, more accurate welfare estimates should be applied.

Future research may look into the transferability of water quality values using tests of in-sample comparisons of CE with TC and CV data. Compared to CV, CEs have been claimed to be more suitable for BT, and function transfer in particular [*Morrison et al*., 2002; *Martin-Ortega et al*., 2012], because they can provide estimates of the WTP for different aspects of a water quality scenario, such as different recreational or nature amenities and site-characteristics [but see *Hanley et al*., 2006a, 2006b]. Such studies may help to better inform decision-makers on the selection of primary studies for BT.

Acknowledgments
This study was carried out as part of the ESRS funded SEER project (RES-060-25-0063, see http://www.cserge.ac.uk) and the EU DG Research funded AquaMoney project (SSPI-022723). We also thank colleagues at CSERGE for valuable help and suggestions and are very grateful to anonymous reviewers for a set of extremely insightful comments on earlier drafts of this paper. Remaining errors are the sole responsibility of the authors.

References

Azavedo, C. D., J. A. Herriges, and C. L. Kling (2003), Combining revealed and stated preferences: Consistency tests and their interpretations, *Am. J. Agric. Econ.*, *85*(3), 525–537, doi:10.1111/1467-8276.00453.

Barton, D. (2002), The transferability of benefit transfer: Contingent valuation of water quality improvements in Costa Rica, *Ecol. Econ.*, *42*(1–2), 147–164, doi:10.1016/S0921-8009(02)00044-7.

Bateman, I. J., et al. (Eds.) (2002), *Economic Valuation with Stated Preference Techniques: A Manual*, Edward Elgar, Cheltenham, U. K.

Bateman, I. J., et al. (2011), Making benefit transfers work: Deriving and testing principles for value transfers for similar and dissimilar sites using a case study of the non-market benefits of water quality improvements across Europe, *Environ. Resour. Econ.*, *50*, 368–387, doi: 10.1007/s10640-011-9476-8.

Bockstael, N., and K. McConnell (Eds.) (2007), *Environmental and Resource Valuation with Revealed Preferences: A Theoretical Guide to Empirical Models*, Springer, Dordrecht, Netherlands.

Boyle, K., M. Welsh, and R. Bishop (1993), The role of question ordering and respondent experience in contingent valuation studies, *J. Environ. Econ. Manage.*, *25*, S80–S99, doi:10.1016/j.bbr.2011.03.031.

Boyle, K. J., and J. C. Bergstrom (1992), Benefit transfer studies: Myths, pragmatism, and idealism, *Water. Resour. Res.*, *28*, 675–683, doi: 10.1029/91WR02591.

Boyle, K. J., N. V. Kuminoff, C. F. Parmeter, and J. C. Pope (2010), The benefit-transfer challenges, *Annu. Rev. Resour. Econ.*, *2*, 161–182, doi: 10.1146/annurev.resource.012809.103933.

Brander, L. M., R. J. G. M. Florax, and J. E. Vermaat (2006), The empirics of wetland valuation: A comprehensive summary and a meta-analysis of the literature, *Environ. Resour. Econ.*, *33*, 223–250, doi:10.1007/s10640-005-3104-4.

Brouwer, R. (2000), Environmental value transfer: State of the art and future prospects, *Ecol. Econ.*, *32*, 137–152, doi:10.1080/09640560802207860.

Brouwer, R. (2008), The potential role of stated preference methods in the Water Framework Directive to assess disproportionate costs, *J. Environ. Plann. Manage.*, *51*(5), 597–614.

Brouwer, R., and I. J. Bateman (2005), Temporal stability and transferability of models of willingness to pay for flood control and wetland conservation, *Water Resour. Res.*, *41*, W03017, doi:10.1029/2004WR003466.

Brown, T. C., P. A. Champ, R. C. Bishop, and D. W. McCollum (1996), Which response format reveals the truth about donations to a public good?, *Land. Econ.*, *72*, 152–166.
Cameron, T. A., G. L. Poe, R. G. Ethier, and W. D. Schulze (2002), Alternative non-market value-elicitation methods: Are the underlying preferences the same?, *J. Environ. Econ. Manage.*, *44*, 391–425, doi:10.1006_jeem.2001.1210.
Cameron, T. A., and D. D. Huppert (1991), Referendum Contingent Valuation Estimates - Sensitivity to the Assignment of Offered Values, *J. of the Amer. Stat. Ass.*, 86(416), 910–918, doi:10.1080/01621459.1991.10475131.
Champ, P. A., K. Boyle, and T. C. Brown (Eds.) (2003), *A Primer on Non-market Valuation, The Economics of Non-Market Goods and Services*, vol. 3, Kluwer Acad., Dordrecht, Netherlands.
Desvousges, W. H., M. C. Naughton, and G. R. Parsons (1992), Benefit transfer: Conceptual problems in estimating water quality benefits using existing studies, *Water. Resour. Res.*, *28*, 675–683, doi:10.1029/91WR02592.
Fezzi, C., S. Ferrini, and I. J. Bateman, (2014), Using revealed preferences to estimate the value of travel time to recreation sites, *J. Environ. Econ. Manage.*, *67*(1), 58–70, doi:10.1016/j.jeem.2013.10.003.
Haab, T. C., and K. E. McConnell (Eds.) (2002), *Valuing Environmental and Natural Resources: The Econometrics of Non-Market Valuation*, Edward Elgar, Cheltenham, U. K.
Hanemann, W. M. (1989), Welfare evaluation in contingent valuation welfare evaluations in contingent valuation experiments with discrete responses data: Reply, *Am. J. Agric. Econ.*, *66*, 332–341, doi:10.2307/1242685.
Hanley, N., R. E. Wright, and B. Alvarez-Farizo (2006a), Estimating the economic value of improvements in river ecology using choice experiments: An application to the Water Framework Directive, *J. Environ. Manage.*, *78*, 183–193, doi:10.1016/j.jenvman.2005.05.001.
Hanley, N., S. Colombo, D. Tinch, A. Black, and A. Aftab (2006b), Estimating the benefits of water quality improvements under the Water Framework Directive: Are benefits transferable?, *Eur. Rev. Agri. Econ.*, *33*(3), 391–413, doi:10.1093/eurrag/jbl019.
Hime, S., I. J. Bateman, P. Posen, and M. Hutchins (2009), A transferable water quality ladder for conveying use and ecological information within public surveys, *CSERGE Working Pap. EDM 09-01*, Univ. of East Anglia, Norwich, U. K.
Hynes, S., N. Hanley, and C. O'Donoghue (2009), Alternative treatments of the cost of time in recreational demand models: An application to white water kayaking in Ireland, *J. Environ. Manage.*, *90*, 1014–1021, doi:10.1016/j.jenvman.2008.03.010.
Johnston, R. J., and R. S. Rosenberger (2010), Methods, trends and controversies in contemporary benefit transfer, *J. Econ. Surv.*, *24*, 479–510, doi:10.1111/j.1467-6419.2009.00592.x.
Johnston, R. J., E. Y. Besedin, and M. H. Ranson (2006), Characterizing the effects of valuation methodology in function-based benefit transfer, *Ecol. Econ.*, *60*, 407–419, doi:10.1016/j.ecolecon.2006.03.020.
Kataria, M., I. Bateman, T. Christensen, A. Dubgaard, B. Hasler, S. Hime, J. Ladenburg, G. Levin, L. Martinsen, and C. Nissen (2012), Scenario realism and welfare estimates in choice experiments: A non-market valuation study on the European water framework directive, *J. Environ. Manage.*, *94*, 25–33.
Kealy, M. J., and R. W. Turner (1993), A test of the equality of closed-ended and open-ended contingent valuations, *Am. J. Agric. Econ.*, *75*, 321–331, doi:10.2307/1242916.
Kirchhoff, S., B. G. Colby, and J. T. LaFrance (1997), Evaluating the performance of benefit transfer: An empirical inquiry, *J. Environ. Econ. Manage.*, *33*, 75–93, doi:10.1006/jeem.1996.0981.
Kristofersson, D., and S. Navrud (2005), Validity tests of benefit transfer: Are we performing the wrong tests?, *Environ. Resour. Econ.*, *30*, 279–286, doi:10.1007/s10640-004-2303-8.
Lew, D. K., and D. M. Larson (2008), Valuing a beach day with a repeated nested logit model of participation, site choice, and stochastic time value, *Mar. Resour. Econ.*, *23*(3), 233–252.
Loomis, J., M. Creel, and T. Park (1991), Comparing benefit estimates from travel cost and contingent valuation using confidence intervals for Hicksian welfare measures, *Appl. Econ.*, *23*, 1725–1731, doi:10.1080/00036849100000067.
Loomis, J., B. Roach, F. Ward, and R. Ready (1995), Testing transferability of recreation demand models across regions: A study of corps of engineer reservoirs, *Water. Resour. Res.*, *31*, 721–730, doi:10.1029/94WR02895.
Loomis, J. B. (1992), The evolution of a more rigorous approach to benefit transfer: Benefit function transfer, *Water. Resour. Res.*, *28*, 701–705, doi:10.1029/91WR02596.
Martin-Ortega, J., R. Brouwer, E. Ojea, and J. Berbel (2012), Benefit transfer of water quality improvements and spatial heterogeneity of preferences, *J. Environ. Manage.*, *106*, 22–29, doi:10.1016/j.jenvman.2012.03.031.
Metcalfe, P. J., et al. (2012), An assessment of nonmarket benefits of the Water Framework Directive for households in England and Wales, *Water. Resour. Res.*, *48*, W03526, doi:10.1029/2010WR009592.
Meyerhoff, J., M. Boeri, and V. Hartje (2014), The value of water quality improvements in the region Berlin–Brandenburg as a function of distance and state residency, *Water Res. Econ.*, doi:10.1016/j.wre.2014.02.001, in press.
Morey, E. R., R. D. Rowe, and M. Watson (1993), A repeated nested-logit model of Atlantic salmon fishing, *Am. J. Agric. Econ.*, *75*, 578–592, doi:10.2307/1243565.
Morrison, M., J. Bennett, R. Blamey, and J. Louviere (2002), Choice modelling and tests of benefit transfer, *Am. J. Agric. Econ.*, *84*, 161–170, doi:10.1111/1467-8276.00250.
Muthke, T., and K. Holm-Mueller (2004), National and international benefit transfer testing with a rigorous test procedure, *Environ. Resour. Econ.*, *29*, 323–336, doi:10.1007/s10640-004-5268-8.
Mwebaze, P., and J. Bennett (2012), Valuing Australian botanic collections: A combined travel cost and contingent valuation approach, *Austr. J. of Agric. and Res. Econ.*, *56*, 498–520, doi: 10.1111/j.1467-8489.2012.00595.x.
Needelman, M. S., and M. J. Kealy (1995), Recreational swimming benefits of New Hampshire lake water quality policies: An application of a repeated discrete choice model, *Agric. Resour. Econ. Rev.*, *24*, 78–87.
Poe, G. L., E. K. Severance-Lossin, and M. P. Welsh (1994), Measuring the Difference (X - Y) of Simulated Distributions: A Convolutions Approach, *Amer. J. Agr. Econ.*, *76*(4), 905–915, doi: 10.2307/1243750.
Ready, R. C., J. C. Buzby, and D. Hu (1996), Differences between continuous and discrete contingent value estimates, *Land. Econ.*, *72*, 397–411.
Rolfe, J. C., and B. Dyack (2010), Testing for convergent validity between travel cost and contingent valuation estimates of recreation values in the Coorong, Australia, *Austr. J. of Agric. and Res. Econ.*, *54*, 583–599 doi: 10.1111/j.1467-8489.2010.00513.x.
RPA (2003), *Water framework directive: Indicative costs of agricultural measures, final report*, Dep. for the Environ. Food and Rural Affairs, Loddon, Norfolk.
Santagata, W., and G. Signorello (2000), Contingent valuation of a cultural public good and policy design: The case of 'Napoli Musei Aperti,' *J. Cult. Econ.*, *24*, 181–204, doi:10.1023/A:1007642231963.
Schaafsma, M., R. Brouwer, and J. Rose (2012), Directional heterogeneity in WTP models for environmental valuation, *Ecol. Econ.*, *79*, 21–31.

Shrestha, R. M., and J. B. Loomis (2003), Meta-analytic benefit transfer of outdoor recreation economic values: Testing out-of-sample convergent validity, *Environ. Resour. Econ.*, *25*, 79–100, doi:10.1023/A:1023658501572.

Smith, V. K., W. Desvousges, and M. McGivney (1983), The opportunity cost of travel time in recreation demand models, *Land. Econ.*, *59*(3), 259–278.

Smith, V. K., W. H. Desvousges, and A. Fisher (1986), A comparison of direct and indirect methods for estimating the environmental benefits, *Am. J. Agric. Econ.*, *68*, 280–291, doi:10.2307/1241429.

Willis, K. G., R. Scarpa, and M. Acutt (2005), Assessing water company customer preferences and willingness to pay for service improvements: A stated choice analysis, *Water Resour. Res.*, *41*, W02019, doi:10.1029/2004WR003277.

Wolff, H., (2014), Value of time: Speeding behavior and gasoline prices, *J. Environ. Econ. Manage.*, *67*(1), 71–88, doi:10.1016/j.jeem.2013.10.002.