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

by

Bateman, I.J., Brouwer, R., Ferrini, S., Schaafsma, M., Barton, D.N., Dubgaard, A., Hasler, B., Hime, S., Liekens, I., Navrud, S., De Nocker, L., Ščeponavičiūtė, R., and Semėnienė, D.

CSERGE Working Paper EDM 09-10

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[1]

Bateman[1*], I.J., Brouwer, R.[2], Ferrini, S.[1], Schaafsma, M.[2], Barton, D.N.[3], Dubgaard, A.[4], Hasler, B.[4], Hime, S.[1], Liekens, I.[5], Navrud, S.[6], De Nocker, L.[5], Ščeponavičiūtė, R.[7], and Semėnienė, D.[7]

[1] School of Environmental Sciences, University of East Anglia, Norwich, UK.
[2] Department of Environmental Economics, Institute for Environmental Studies, VU University, Amsterdam, Netherlands.
[3] Norwegian Institute for Nature Research, Oslo, Norway.
[4] Institute of Food and Resource Economics, University of Copenhagen, Denmark.
[5] Flemish Institute for Technological Research, Boeretang, Belgium.
[6] Department of Economics and Social Sciences, Agricultural University of Norway, Aas, Norway.
[7] Center for Environmental Policy, Vilnius, Lithuania.

Abstract

We develop and test guidance principles for benefits transfers. These argue that when transferring across relatively similar sites, simple mean value transfers are to be preferred but that when sites are relatively dissimilar then value function transfers will yield lower errors. The paper also provides guidance on the appropriate specification of transferable value functions arguing that these should be developed from theoretical rather than ad-hoc statistical principles. These principles are tested via a common format valuation study of water quality improvements across five countries. Results support our various hypotheses providing a set of principles for future transfer studies. The application also considers new ways of incorporating distance decay, substitution and framing effects within transfers and presents a novel water quality ladder.

Keywords
Non-market valuation; stated preference; benefit transfers; transfer errors; methodology; water quality

JEL classifications
Q51, Q15, Q26, Q24, Q28

[*] Corresponding author. E-mail: i.bateman@uea.ac.uk.

1. Introduction

Decision making is an essential yet costly undertaking and resource constraints inevitably mean that that the decision process itself has to pass cost-benefit tests. Analysts have for many years sought methods which will reduce decision costs, and the extrapolation of assessments from one case to another is clearly attractive. Given the significant costs of valuing preferences for non-market goods, it is not surprising that this area has now generated a considerable literature concerning the transfer of value estimates, most particularly in the area of environmental valuation (Brouwer, 2000). Such transfer exercises typically involve estimating the value of a given change in provision of a good at some target 'policy site' from analyses undertaken previously at one or more 'study site'. The most fundamental problem for value transfers is in assessing whether a given transfer is correct or not when the 'true' value of the policy site is a-priori unknown.

The literature has placed great emphasis upon the development and testing of value transfer methods (e.g., Desvousges et al., 1992; Bergland et al., 1995; Downing and Ozuna, 1996; Brouwer and Spaninks, 1999; Ready et al., 2004; Zandersen et al 2007; Johnston and Duke 2009)[2]. These methods can be broadly categorised into two types (Navrud and Ready, 2007a). The simplest approach is to transfer mean values from study to policy sites (e.g. Muthke and Holm-Mueller, 2004). Such transfers are frequently used in practical decision making but are crucially dependent upon the pertinence of differences between transfer sites. Clearly at some level all sites are dissimilar (e.g. the unique ecosystem habitats or the spatial pattern of substitutes around a site are unique); it is the degree to which this dissimilarity affects values which will determine the appropriateness of such 'univariate transfers'. The principal alternative is to use statistical techniques to estimate value functions from study site data. These are then used to predict new values for policy sites. This multivariate 'value function transfer' approach assumes that the underlying utility relationship embodied in the parameters of the estimated model applies not only to individuals at the study sites but also to those at policy sites. While parameters are kept constant, the values of the explanatory variables to which they apply are allowed to vary in line with the conditions at the policy site.

One of the major objectives of this paper is to develop and test simple rules for conducting benefit transfers based upon the hypothesis that univariate approaches will be more appropriate for transferring between relatively similar sites, while transferable value functions will yield lower errors for transfers between less similar sites. The intuition behind such a hypothesis is also straightforward with the same driver responsible for these different outcomes. Value functions explicitly incorporate differences between sites. Where these differences are relatively extreme a *well specified* value function will reflect and incorporate this heterogeneity, so providing a better estimate of the value of a policy site than is afforded by a simple univariate mean transfer. However, when transferring between similar sites the incorporation of differences inherent in the value function approach may generate higher degrees of error than the simple transfer of mean values between those relatively homogeneous sites.

An immediate question which such a hypothesis generates is how an analyst would determine (a-priori and without survey evidence from policy sites) whether sites where similar or not. We assess this through a simple examination of secondary source data regarding the characteristics of sites and their surrounding populations, examining whether

---

[2] For ease of exposition we omit discussion of parallel approaches such as meta analysis (e.g. Bateman and Jones, 2003; Lindhjem and Navrud, 2008) and Bayesian approaches to modelling value functions (e.g. Moeltner et al., 2007, Leon-Gonzalez and Scarpa, 2008).

such information is sufficient to determine similarity and hence the appropriate choice of transfer method.

In essence our study tries to make sense of the conflicting evidence available in the benefit transfers literature, wherein some studies report higher errors from univariate transfers than value function approaches while others find the opposite result. (see, for example, Bergland et al., 1995; Barton, 2002; Chattopadhyay, 2003; Ready et al., 2004; Brouwer and Bateman, 2005). However, the paper has further related objectives. We argue that an additional reason for these diverse results is misspecification of empirical value functions for transfer purposes. We do not claim that the value functions used previously are statistically erroneous, but rather argue that what is appropriate for statistical specification (maximisation of explained variance) may be inappropriate for function transfer purposes (minimisation of transfer error). We contend that many value function transfer exercises have failed because they have employed ad-hoc, empirically driven specification of utility rather than drawing upon the common drivers of preferences reflected in economic theory. Theory suggests that the utility of improvements to a spatially confined, environmental resource should be determined by the change in provision, characteristics of the site (e.g. its distance from the valuing individual), the availability of substitutes (e.g., distance to substitutes) and characteristics of the valuing individual common to all utility functions (e.g., the individual's income). Of course the value of improvements at a given study site may also be influenced by context specific factors unique to (or of particular relevance to) that site. While the inclusion of such contextual variables may improve the degree to which a value function explains values at the study site, by including such factors within a transferred function we assume that the relevance of that variable holds for the policy site. To the extent that this is not the case, so the value function transfer will generate error. Because of the multiplicative nature of a value function it seems clear that such error has the potential to be substantial. We test the contention that value functions specified to only include those variables which economic theory expects to be common to all contexts will generate lower transfer errors than functions which include ad-hoc (non-theoretic) variables, even when the latter functions provide higher degrees of statistical fit at the study site.

The paper also offers a number of further contributions. First, as part of the specification of economic-theoretic value functions we utilise geographic information systems (GIS) to assess the impact of the location of both improvement and substitute sites upon values. The measures obtained reveal decay in values as the distance between improvement site and the survey respondent's home increases. In contrast stated values rise as the distance to substitutes increases. Both relationships conform directly with prior economic-theoretic expectations yet, to our knowledge, this is the first time that such patently important variables have been included within benefit transfer analyses. Second, we consider the challenge of behavioural economics (Camerer, et al., 2003; Douglas and Rangel, 2008; Mullainathan and Thaler, 2001; Rabin, 1998; Simon, 1987) by developing an approach to the assessment of scope sensitivity which addresses the potential for framing effects (DeShazo, 2002; Slovic, 1995). These effects are observed when changes in the way in which objectively identical valuation questions are framed yield significant differences in the responses given by survey respondents. Such anomalies have frequently been observed in the valuation literature (see review in Bateman et al., 2002a) but again, to our knowledge, have not previously been allowed for in benefit transfer analyses. A final additional contribution is the introduction of a novel tool for conveying changes in the quality of the environmental good under consideration (water quality). Building upon ecological assessments of the impacts of diffuse water pollution upon aquatic life and general ecosystem quality, we develop a new water quality ladder incorporating information pertinent to both use and non-use values and presented in a readily comprehended visual form.

The remainder of this paper is organized as follows. The next section provides further detail regarding the motivation for our focus upon site similarity when choosing transfer methodology. This section also clarifies the key principles and expectations which theory provides for undertaking and assessing valid and robust value function transfers. Section 3 discusses our case study test of the arguments summarised above. This concerns a common design, contingent valuation (CV) assessment of willingness to pay (WTP) for improvements in the quality of open waters in Europe; a topic chosen because of its current policy relevance given the incipient implementation of the EU Water Framework Directive (WFD; European Parliament, 2000) which mandates improvements in all European rivers. This section also introduces the case study areas which are distributed across five European countries chosen to include both relatively similar and dissimilar sites. Test procedures are specified to examine our central arguments. As part of this discussion we also consider the importance of allowing for distance decay, substitution and framing effects within valuation studies and overview a new water quality ladder for conveying information on water quality to survey respondents. Section 4, reports results. This opens with a consideration of measures of case study similarity and reports findings supporting the central contention that, when sites are relatively similar, simple mean value transfers minimise transfer errors but that value function methods are required when transferring across dissimilar sites. Results also show that, in the latter case, value function transfer errors are lower when those functions are specified to only include those generic drivers of utility highlighted by economic theory rather than transferring ad-hoc statistical best-fit functions. Section 5 concludes with a discussion of the general implications arising from this research.

2. Expectations based principles for benefit transfers

Pearce et al. (1994) argue that, in principle, because value function transfers allow the analyst greater control over differences across sites, they should yield lower transfer errors than simple mean value transfers.

In essence the present paper adds two qualifiers to that argument. First, this may not hold for transfers across relatively similar sites. Mean values smooth out the variation which inevitably arises when a sample survey is undertaken. Value functions will, in comparative terms, give a greater reflection of the variability of a sample. Estimated coefficients can be influenced by small numbers of individuals in a sample, especially where they have relatively extreme characteristics or unusual values. Because samples from two (or more) sites are likely to contain different proportions of such (relatively) extreme observations (even when those sites and their surrounding populations are relatively similar), when estimated coefficients are used within a value function transfer they may yield considerably greater errors than simple mean value transfers. The value function approach assumes that the relationships reflected in estimated coefficients is constant across sites. Even if this is the underlying truth, the vagaries of any sampling exercise may yield variations in estimated coefficients which result in higher transfer errors when the value function approach is used than when simple mean value transfers are applied.

While the above argument seems likely to apply for similar sites, the opposite seems likely to apply for dissimilar sites. Here the differences between the values of the underlying population are likely to be so gross that simple mean transfers will yield substantial errors. However, now the value function approach comes into its own. Accepting that the sampling variation effects described above will still apply, nevertheless the value function method now has the ability to adjust for the gross differences between dissimilar sites; be they physical or socioeconomic/demographic in nature. Simple mean value transfers will be unable to make such adjustments here and in principle are liable to yield larger errors than value function approaches. However, here we encounter our second caveat to the Pearce et al., argument;

that value functions need to be specified for the purposes of transferral between study and policy sites rather than for obtaining statistical best-fit at study sites alone.

In specifying value functions for transfer purposes we have to ensure that the predictors used are of generic relevance to both study and policy sites. This rules out the inclusion of context specific ad-hoc variables. While such predictors may significantly assist in optimising the statistical fit of models to study site data, the danger is that they will be of different (or even no) relevance to policy sites. In effect the assumption that parameter estimates from study sites will hold for policy sites begins to fail. Due to the multiplicative role of coefficients, such failure can result in major transfer errors. Our contention is that statistical best-fit guidelines for function specification have to be abandoned when the purpose of that model building is for transfers. Instead the focus should be restricted to generic factors. We argue that economic theory provides us with a number of expected relationships which should hold across contexts and it is these factors which should guide the specification of transferable value functions.

So what general relationships does economic theory suggest should hold across contexts? While theory leaves much to the discretion of individuals, basic microeconomics principles identify a number of factors which should influence preferences and WTP. Within the context of the a spatially fixed water quality public good considered in our case study, these factors include: the extent (or 'scope') of the change in provision under consideration; the costs which an individual faces for using the good (for an open-access good this mainly relates to the proximity of the good to the respondent's home); the availability of substitutes (again a spatial relationship) and the individual's income constraints. Theory also implies that responses should be invariant to the procedure used to elicit WTP providing that the questions being asked are objectively identical. These factors provide a rich source of theoretically consistent variables for inclusion in our theory driven value functions and we consider each in turn within the remainder of this section. Furthermore, their inclusion within our value functions means that we have a number of clear expectations against which we can validate our findings.

The US NOAA Blue-Ribbon Panel (Arrow et al., 1993) highlighted the responsiveness of WTP responses to changes in provision as the principal form of validity assessment for CV studies. This is based upon the expectation that:

> "[u]sually, though not always, it is reasonable to suppose that more of something regarded as good is better so long as an individual is not satiated. This is in general translated into a willingness to pay somewhat more for more of a good, as judged by the individual" (Arrow et al., 1993; p. 4604).

This 'scope sensitivity' assessment has come to be "regarded by many as an acid test" (Carson et al., 1996, p. 3) of survey-derived values. However, as Banerjee and Murphy (2005) point out, statistically significant sensitivity to scope is of itself an insufficient test of preference consistency. There are very few non-market goods for which we have prior expectations regarding the degree of increase in WTP might be reasonable. Indeed, given that individuals may become satiated with environmental goods (e.g., it might be reasonable for a respondent to think that once they had access to one nearby clean river they were not willing to pay anything for a second), then the only definite expectation that economic theory provides for us is that marginal WTP should not be negative for an increase in provision of a good. Therefore, while we design our case study to allow us to include changes in provision within the transferable value function we consider a finding of significant scope sensitivity as a necessary but insufficient test of study validity[3].

Arguably some of the clearest expectations arise from the spatially fixed nature of environmental public goods (Zandersen et al., 2007) such as open-access water quality. Theory suggests that usage and net benefits will be related to the travel costs faced by households and indeed WTP values tend to decline markedly as the distance to the site in question increases. This trend in average WTP reflects a fall in the proportion of users to non-users as distance increases (Bateman et al., 2006). As users typically hold higher values than non-users a 'distance decay' (*ibid.*) in values is to be expected. This can readily be tested for by recording the home address of survey respondents and using GIS or similar software to assess distance and/or travel times. Analysis of empirical distance decay trends in values can then be undertaken. We therefore include such variables within our transferable value function.

Such accessibility measures can also be used to test a further clear expectation regarding location; that WTP for an improvement at a given site should decline as the availability of suitable substitutes rises. An operational issue here concerns the definition of substitutes. Reliance upon self-reported substitutes involves some challenging questions for survey respondents and generates variables which are not available to the decision maker wanting to estimate values for unsurveyed sites. Therefore we follow the approach of Jones et al., (2002) in using a GIS to calculate distances to multiple potential substitutes. The latter study then allows the data to determine which are the significant substitutes which is an appealing strategy. However, findings show, not surprisingly, that it is the nearest similar substitute which has the greatest influence. Given this result we adopt a simple approach by including the distance to the nearest similar substitute within our transferable value function.

While a variety of socio-economic and demographic variables may empirically influence stated values, theoretical expectations emphasise the role of income in terms of the budget constraints it may impose on WTP. We might expect, *ceteris paribus*, that those with higher incomes will have higher WTP. In a manner analogous to the scope test, this is a fairly weak expectation dependent upon the value of the good in question. Nevertheless, income effects do appear to have some microeconomic foundations and are included in our transferable value functions.

While microeconomic practice provides some expectations regarding what factors should influence values, theory also indicates certain issues which should <u>not</u> affect WTP and this provides a last set of design concerns for our transferable value functions. In essence, economic theory posits that, prior to giving a valid WTP response, individuals should have well formed preferences, conforming to standard assumptions and robust against what theory would see as irrelevant issues, such as the way in which a given question was framed. Tests of such procedural invariance are therefore important ways of validating stated preference responses. However, findings from such tests are not uniformly supportive with some results suggesting that individuals may determine their assessments of certain goods not solely by reference to what might be recognised as their economic preferences, but also by inferring information from the manner in which a question is framed (Lichtenstein

[3] Findings of significant scope sensitivity are more convincing for cross-respondent valuations obtained split sample CV studies (where different groups of respondents are presented with different scopes of the good in question) than from within-respondent values (where the same respondent reacts to different scopes of the good. This is because the latter can be criticised as potentially merely reflecting a respondent's desire for internal coherence in their responses (Ariely et al., 2003). This internal coherence argument weakens the importance of scope sensitivity tests for choice modelling studies where respondents choose between multiple options, each defined by varying scope of the goods in question.

and Slovic, 1971; Tversky and Kahneman, 1974; Kahneman et al., 1982; Slovic, 1995; Hsee, 1996; Birnbaum, 1999; Doyle et al., 1999; Ariely et al., 2003; Kahneman, 2003; Bateman et al., 1997a,b, 2005, 2007). The problem with such 'constructed' preferences is that they are highly malleable; changing with the frame within which a question is posed. Such phenomena have excited considerable interest within the valuation literature (Kahneman and Knetch, 1992; Schkade and Payne, 1994; Bateman et al., 2008, forthcoming) with commentators arguing that framing effects have to be addressed within the design of studies (DeShazo, 2002; Bateman et al., 2009). We therefore design our study to allow the incorporation of a procedural invariance test within our transferable value function in a manner described in the next section.

Our assertion then is that a value function specified to only include generic, theory derived variables will yield lower transfer errors than a function which improves its statistical fit to the survey site data by including ad-hoc, potentially context specific, variables. To provide a Popperian falsifiable test of this hypothesis we also examine the influence upon valuation transfers that inclusion of variables not suggested by economic expectations may have. Clearly we could identify extremely ad-hoc variables that could not reasonably apply to more than one or two study sites. However, such an artificial test would not validate out argument. Instead we include a small number of variables which have appeared as significant predictors of values within the literature. If our central hypothesis holds then the inclusion of such 'empirical regularity' variables should increase transfer errors over those associated with functions specified purely upon economic expectations.

3. Developing and implementing the common design

The study design followed a set of valuation design principles set out in Bateman et al., 2002, and the Appendix to this paper). Initial concerns for study design were to identify a public good and case study locations to provide a rigorous yet policy relevant test of our methodology. Considering the latter locational issue, recall that the underlying objective of value transfer is simple; to take information on the value of provision changes at some surveyed study site(s) and with it estimate values for provision changes at some unsurveyed policy site(s). However, we first need to be sure that the transfer methods employed are valid and reliable. To achieve this requires survey data from at very least two sites. Transfer then involves using data from, say, site A to predict values at site B. Validation then compares the value of site B as predicted by transfers from site A with the actual value obtained from the survey of site B (with the transfer error being expressed in terms of the percentage difference between the two WTP estimates; see, for example, Bergland et al., 1995). So, while the objective is to develop methods for transferring to unsurveyed sites, methodological development requires data at all sites. To test out our central hypotheses regarding the importance of site similarity in selecting the most appropriate method for transfers we sought to select case study sites from both similar and dissimilar contexts. With this objective in mind, research collaboration was agreed between five European countries: Belgium, Denmark, Lithuania, Norway and the UK. As detailed in subsequent results, together these include both similar and dissimilar countries capturing some of the economic extremes of Europe and providing a robust test for our methodology.

While we deliberately sought variation in the study site contexts, we need to value a common good in all cases. It was felt that this good should be typical of those assessed within non-market valuation studies; one which generates both use and non-use values, of relevance across all case study areas and of policy interest. As can be seen by the number of meta-analyses of surface water (e.g., Johnston et al., 2005, 2006; Moeltner et al., 2007), this is a common target for non-market valuation that has from the earliest of studies reported significant use and non-use values arising from water quality improvements (e.g.

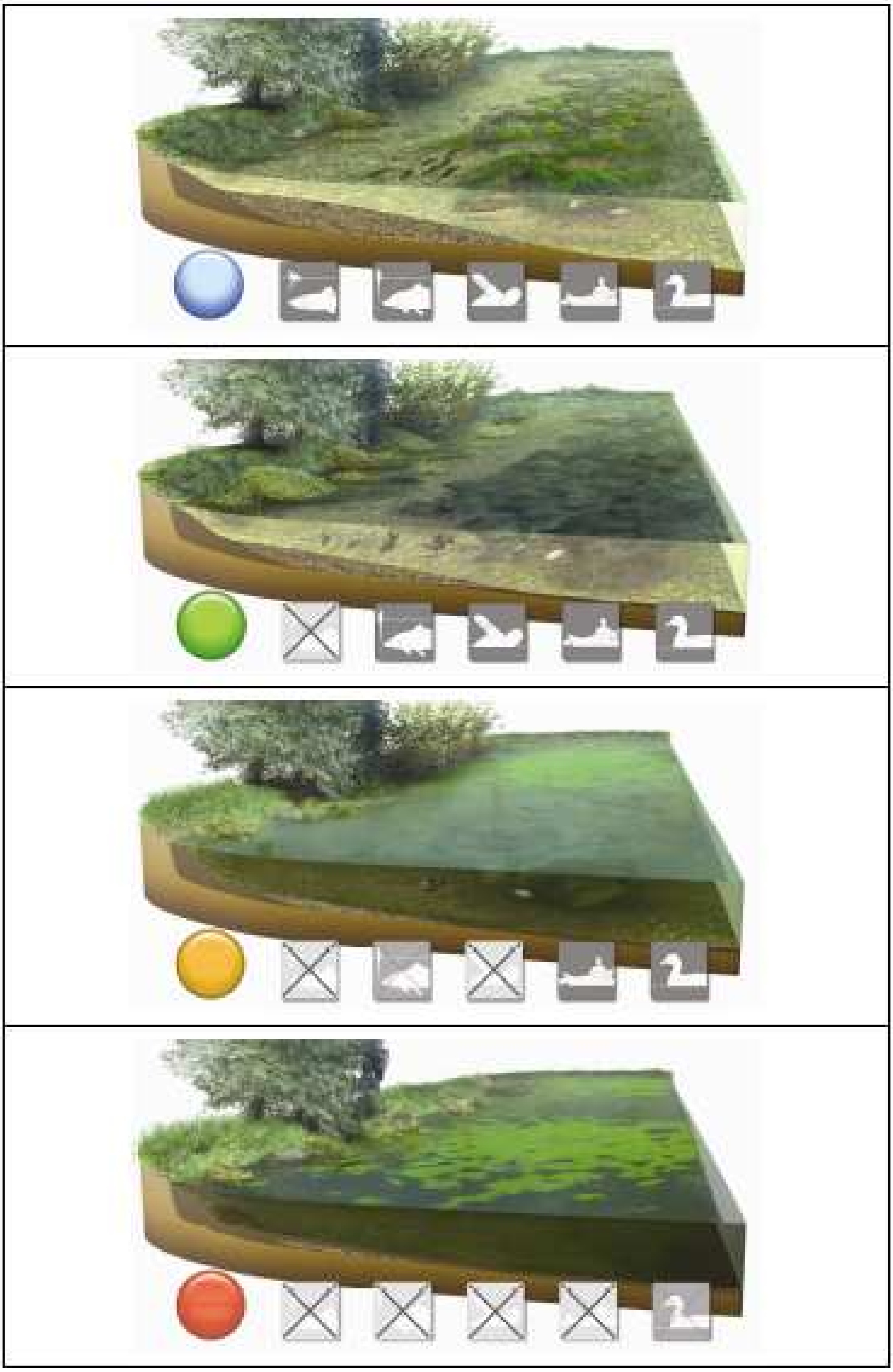
Desvousges et al., 1987). Furthermore, this literature supports the common sense notion that open-access water quality is of interest to almost all populations, allowing us to undertake studies in multiple countries and transfer between them. Finally, with the introduction and gradual ongoing implementation of the WFD this is a topic of great policy interest. The WFD represents a fundamental change in the management of water quality in Europe with a general requirement to improve all European waters to "good ecological status" by 2015. In the five northern European case study areas the main water quality problem is eutrophication[4]. Moreover, there is a common policy need for information to justify time derogations and the setting of less restrictive targets in cases of disproportionate costs as determined through economic assessment of costs and benefits (WATECO, 2004). These issues provide a common ground to the valuation scenario.

A vital early task in any SP valuation study is the clear definition of the good concerned, its status quo conditions and the change(s) in provision which we will ask survey respondents to value. This in turn requires an understanding of the physical science determining these states. While there are numerous pollutants that affect open access waters, the WFD focuses upon those which affect their ecological status and in particular those nutrients that are delivered to waterways via routes such as diffuse pollution from agriculture (Davies and Neal, 2004; Hutchins et al., 2006). To some considerable degree the pathways linking pollution to ecological impact is still the subject of ongoing research (UKTAG, 2008). However, this does not prevent the analyst from valuing certain states of the world on the assumption that ongoing research will indicate how such states might subsequently be attained. Furthermore, individuals do not hold values for reducing pollution per se but rather for the effects that such reductions may induce in terms of recreation suitability, ecological quality, etc.

Clear and comprehensible information is essential for ensuring understanding of a good and its provision changes within a stated preference survey. The extensive literature on information provision stresses the advantages of visual as opposed to textual or numeric approaches (e.g., Peters et al., 2005; Fagerlin et al., 2005; Zikmund-Fisher et al., 2005; Bateman et al., 2009) and this is reflected within the water quality valuation literature (e.g., Carson and Mitchell, 1993). With this in mind, a novel 'water quality ladder' was developed for the present application (full details being given in Hime et al., 2009 with a summary provided in the Appendix to this paper). This defined four levels of water quality based upon chemical, physical, flora and fauna characteristics as well as use characteristics. Following discussions across the various case study partners, a set of photographs of generic water quality characteristics was agreed for each quality level. These were then passed to a graphic artist to produce the generic water quality ladder shown in Figure 1. This ties together the ecological and use attributes of water bodies to be applicable to a wide range of lowland slow flowing rivers as well as lakeshores. The simple colour coding scheme shown in the figure allowed clear definition of quality levels in the survey interview. Qualitative face-to-face testing with a pilot sample confirmed that this form of information was clearly comprehended by respondents who were able to recall patterns in quality change following the interview process.

[4] This contrasts with southern Europe where an increasingly serious problem is water scarcity.

Figure 1: The generic water quality ladder



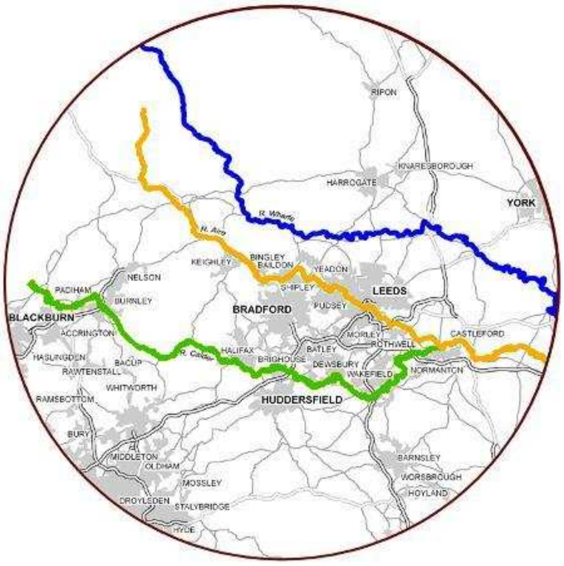
Source: adapted from Hime et al., (2009): Copyright protected.

With the nature of the good clarified, the next task was to define the current level of provision and changes in that provision, which together determine scope. For rigorous scope sensitivity testing we require a clear definition of the status quo and at least two changes in provision of the good (a single provision change does not allow us to examine the shape of a value function or assess changes in the marginal value of a resource as its provision alters). These changes in provision need to be defined in terms of both quality and quantity. To enhance the consistency of our design, in each country the case study was applied to a water body whose status quo quality was best described by the yellow level of the water quality ladder while the quality of improved stretches was specified as attaining the blue level[5]. The two changes in provision were then distinguished by defining a waterbody improvement (which we will term the Large improvement) and halving this to produce a second scope of change (which we term the Small improvement). The contrast between the values for these two quantities provides an insight into the rate at which marginal WTP diminishes as the scope of improvements increases. Such information is vital to prevent overestimation of values when considering more major improvements than those considered here. Each provision change was presented to respondents in map form with Figure 2 illustrating maps from the UK case study.

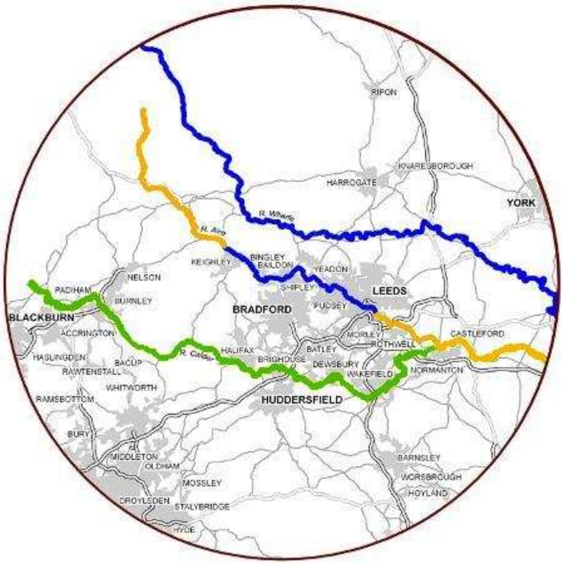
[5] Note that a further treatment in the Norwegian study also examined an improvement from ‘red’ to ‘green’ quality. However, for comparability, this dataset was not used within the present analysis.

Figure 2: Maps from the UK case study depicting status quo provision of river water quality and smaller and larger improvements.

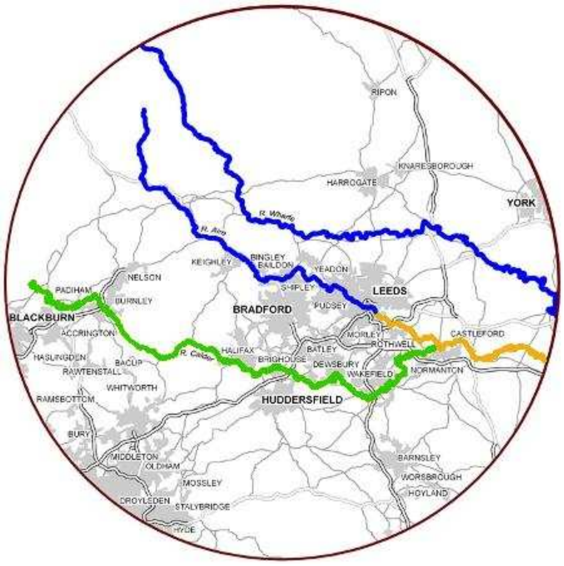
Status quo



Smaller improvement



Larger improvement



Scope sensitivity and diminishing marginal WTP testing involves examination of valuation changes between the Small and Large improvement. There are a number of ways in which the data required for such an assessment can be gathered. One simple route is to ask each respondent to state their WTP for both the Small and Large improvement. As noted previously, such within-respondent scope tests are fairly weak and susceptible to criticism. Clearly an across-respondent test becomes feasible if a split sample approach is adopted where some individuals face an initial valuation question concerning the Small improvement while others face the Large improvement.

A split sample approach also readily facilitates a test of procedural invariance. Using an 'exclusive list' question format (Bateman et al., 2004), immediately after providing their initial WTP answer respondents are asked to imagine that the Small (or Large; depending on question ordering) provision change they had just considered had not occurred and that they were still at their status quo point. Now a second valuation question is asked eliciting WTP for the Large (or Small) change in provision. By ensuring both valuations are made from the common status quo level we avoid the sequencing problems highlighted by Carson et al. (1998) which highlighted the non-comparability of values for a given good made from different baselines. Repeating such a procedure across the sample elicits four values: WTP responses to the first and second questions concerning either the Small or Large improvement (yielding values denoted 'Small improvement 1st'; 'Large improvement 1st'; 'Small improvement 2nd' and 'Large improvement 2nd'). The procedural invariance expectation is therefore that values for the Small improvement should be invariant to the order in which they were elicited, as should be the values for the Large improvement.

Given that the valuation literature provides clear evidence regarding the potential for changes in the WTP elicitation method having significant impacts upon responses (e.g. Bateman et al., 1995) a common approach was used in all countries. This consisted of a payment card presented in local currency units but which, when converted into Euros, included the same amounts for all countries[6]. The payment card amounts were chosen after considering the differences in purchasing power between countries and the impact upon the statistical efficiency of WTP estimates of different payment card levels (see the Appendix to this paper for details including the full text of the valuation question)[7]. The WTP question was prefixed by a standard budget constraint reminder

The common questionnaire also contained uniform questions regarding respondents' household income for incorporation within our transferable value function. Given that we wish to test a theoretically specified function against one derived from statistical principles, the questionnaire also included a variety of other socio-economic and demographic characteristics such as respondent age[8]. While theory is mute regarding the influence of such variables, they are commonly included in valuation functions and yield some empirical regularities. Arguably such variables, if they are genuine regularities reflecting preference relations which hold universally, could usefully be included within transferable value functions. However, our concern is that once we abandon the parsimonious guidance of economic theory there is no clear demarcation of which effects are likely to be common to all areas. In effect we stray toward the ad-hoc inclusion of context specific variables which our central argument repudiates. Therefore, while in our subsequent empirical analysis we examine the effect of including empirical regularities (providing a falsifiable test of the hypothesis that transferable value functions should be limited to variables for which we have economic expectations), the central thrust of our argument would be to exclude such variables from the specification of value function for transfer purposes.

To ensure that the data contained a sufficient level of variation in terms of the distance to the improvement site and to substitutes an efficient spatial sampling strategy was developed. In essence this strategy considered a regular grid of potential interview locations around the study site, assessing each location in terns of its distance (to site and substitutes),

---

[6] The UK case study additionally employed a dichotomous choice WTP elicitation approach with a separate sample of respondents. This data is not considered within the poresent analysis.

[7] A concern with payment cards is that they may be subject to range bias (Covey et al., 2007) where respondents infer that values in the centre of a range are somehow 'correct'. Following the findings of Rowe et al., (1996) we address this by using a payment card with values which ranged from zero to amounts that were clearly implausibly high and therefore not to be construed as having any information value. We also eschew the common habit of using a logarithmic style card with increasingly wide differences between values at the upper end of the range. Again this may be construed as suggesting such values are less plausible. Instead a card using evenly spaced amounts was used. For details see the Appendix to this paper.

[8] Additional questions concerned usage of water and other outdoor recreation resources, respondents' motivations for their WTP response, etc. Only minor variations of procedure were allowed for across studies. For example, in Lithuania, respondents faced prior valuation questions regarding changes in the hydromorphology of the case study (Neris) river. Arguably these may have impacts upon the valuation responses reported in the present paper.

socioeconomic and population density characteristics. Survey locations were chosen such that a full range of data in each of these dimensions was captured. The home address of each survey respondent was recorded and GIS routines were subsequently employed to calculate individual specific distances and travel times to the improvement and all substitute sites.

Sample sizes were designed to support not only conventional parametric validity testing but also cross sub-sample analyses of the procedural invariance tests. In Belgium, Denmark and Norway, the surveys were conducted online, through a marketing company in the two latter countries. In Lithuania and the UK surveys were conducted using face-to-face interviews. Response rates ranged from 12% in the online Belgian survey up to 55% in the Lithuanian face-to-face survey. All surveys were undertaken during 2008.

4. Results

4.1 Descriptive statistics and the (dis)similarity of study sites

All surveys were undertaken during 2008[9] and a total of 3,589 questionnaires were completed across our five study countries[10]. Response options and data coding was common across studies and monetary variables were PPP adjusted and data pooled into a single analysis. Table 1 presents descriptive statistics of each sample together with WTP sums disaggregated by size of provision change and ordering of question presentation.

Table 1. Descriptive statistics and WTP by country

	Lithuania	Belgium	Denmark	Norway	UK	Total
Sample size						
Number of respondents	500	768	754	1133	434	3589
Respondent characteristics						
Mean distance to the improved site (km)	20	21	30	22	10	22
Mean distance to nearest substitute site (km)	1	3	24	7	5	9
Mean annual pre-tax household income tax; € PPP (s.d. in brackets)	9531 (7823)	40877 (19002)	34854 (17708)	24884 (11452)	26686 (16709)	28310 (17730)
Mean Age	48	45	50	45	50	47
Urban (% urban)	63%	45%	79%	41%	78%	58%
Gender (% women)	49%	36%	44%	48%	46%	45%
WTP values in € PPP (standard deviation in brackets)						
Protest bids (% of country sample)	8%	5%	2%	12%	2%	7%

[9] In Belgium, Denmark and Norway, the surveys were conducted online, through a marketing company in the two latter countries. In Lithuania and the UK surveys were conducted using face-to-face interviews. Response rates ranged from 12% in the online Belgian survey up to 55% in the Lithuanian face-to-face survey.

[10] The sample size proved sufficient to support not only the conventional parametric validity testing reported here but also cross sub-sample analyses of the procedural invariance tests reported in the Appendix to this paper.

Average WTP- Small	6 (23)	47 (66)	25 (38)	42 (82)	19 (29)	31 (61)
Average WTP- Large	8 (38)	48 (70)	36 (52)	47 (86)	26 (35)	37 (66)
Average WTP- Small improvement 1st	6 (15)	50 (67)	29 (42)	45 (88)	22 (32)	34 (64)
Average WTP- Large improvement 1st	10 (52)	49 (70)	31 (41)	45 (78)	25 (32)	36 (64)
Average WTP- Small improvement 2nd	6 (29)	43 (66)	21 (33)	38 (76)	16 (25)	28 (57)
Average WTP- Large improvement 2nd	7 (15)	47 (69)	40 (61)	48 (93)	26 (38)	37 (69)

Notes: Income and WTP recalculated based on Purchasing Power Parity indices (World Bank, 2008). Protest bids are excluded in the estimation of WTP descriptive statistics.

The wider representativeness of the final sample is satisfactory with most sample descriptors differing by less than 5% from national statistics. The section of Table 1 headed 'Respondent characteristics' shows those variables which could be obtained from secondary sources such as the census (for socioeconomic and demographic variables) and open source GIS data[11] (for the physical characteristic descriptors) to allow assessments of similarity for unsurveyed policy sites (although given the representativeness of our samples we use report their values for convenience). The first two rows of this section consider distance from respondents home address to the improvement site and their nearest substitute site. Neither of these statistics suggest any clear dissimilarities across study sites with the distance to improvement site in the range households typically travel for recreation and all countries showing that on average respondents have a substitute site closer to them than the improvement site (both factors being patently important determinants of values yet typically ignored in transfer studies). However, the following row shows that there is one major source of dissimilarity between our study countries. Tests confirm that PPP-adjusted household income is substantially lower in Lithuania, being roughly one quarter to one third of the level in the other countries sampled. Our strong, theoretically derived, expectation is that this major dissimilarity would result in a significant difference in stated WTP between Lithuania and other countries. Following our central hypothesis we therefore expect that mean value transfers will generate higher errors than value function transfers when applied across the full set of sites. However, income differences are insignificant across the remaining four countries. Therefore, again following our hypothesis, if we were to omit Lithuania then we would now expect that mean value transfers would outperform value function transfers for the four remaining similar countries. Subsequently we formally test both of these hypotheses.

The remaining rows of this section of Table 1 detail various other sample characteristics which, although not highlighted by theory as determinants of WTP, have been used by analysts seeking ad-hoc variables to improve the statistical fit of study site value functions. None of these variables suggest any further major dissimilarities across countries. We incorporate such factors within the subsequent test of our hypothesis that models containing such ad-hoc variables, while providing a better fit to study site data, may yield higher transfer errors than functions specified solely from theoretically derived, generic predictors.

[11] See Bateman et al., (2002b) for a review of such sources for use witrhin valuation studies.

The final section of Table 1 overviews our WTP valuation results. Here the first row details protest rates identified using the guidelines in Bateman et al. (2002a). These are consistently within the bounds of acceptability suggested by the literature (Mitchell and Carson, 1989; Champ et al., 2003) and pure protestors were excluded from further analyses[12].

Following guidelines for international value transfers (Navrud and Ready, 2007b), WTP responses (and income data) were corrected for differences in purchasing power between countries using indices from the World Development Indicators (World Bank, 2008) and then converted into 2008 Euros. Even a cursory inspection of these results strongly suggests that our classification of Lithuania as dissimilar to the other countries is reflected in WTP values. For both the Small and Large improvements Lithuanian WTP is less than one-third that of the lowest mean given in any of the other countries. This proportion directly echoes the difference in incomes noted above. Recall that, in a real world benefit transfer we would not a-priori have values for the policy sites. What is clear here is that the clear dissimilarities flagged up by variables which can be obtained from secondary sources (the 'Respondent characteristics' discussed previously) do seem to provide relevant indicators of when simple mean value transfers can or cannot be relied upon. Subsequently we formally test these inferences.

A further point to note in Table 1 is clear evidence of diminishing marginal WTP. Recall that the Large improvement provides double the length of highest quality river than the Small improvement. It does appear that in general the former is accorded a higher value but it is clearly note double the latter. As discussed earlier in this paper, this is not of itself an anomalous result as it is perfectly feasible that respondents may have a rapidly diminishing marginal WTP for additional improvements once an initial length of high quality river has been provided. However, a final point to note in Table 1 is some evidence of a failure of procedural invariance in the form of an ordering effect in valuation responses. While there is no particular ordering pattern within the Large improvement values, four countries yield Small Improvement 1st values which are higher than their respective Small Improvement 2nd WTP with the remaining country giving the same value for both. This finding echoes that of Bateman et al., (2004) who find that the value of small goods is elevated when elicited first in an ordering. Again formal parametric tests of these trends are given below (with nonparametric confirmation of these findings detailed in the Appendix to this paper).

4.2 Specification and estimation of value functions

In order to compare simple mean value transfers with value function transfers we first need to estimate the latter functions. To test our assertions regarding the importance of correct specification of transferable value functions we develop a 'theory-driven' model from economic principles. This includes all those variables for which economic theory holds expectations and which should be generic to all sites: the change in provision; the costs of using the good (its distance from the valuing individual); the availability of substitutes (distance to substitutes) and budget constraints (the household income). Because income is the main dimension of dissimilarity between sites we also estimate a further model excluding

[12] We retain the 4% of WTP responses clash with prior expectations in that the smaller improvement is accorded a higher value than the larger good. While this is an issue to be highlighted (and is often not tested for in non-market valuation studies) the rate of apparent irrationality or misunderstanding of the scenario is consistent with findings in experimental economic tests. While some studies have omitted data from such respondents we argue that this may give a misleading indication of the consistency and validity of findings and so retain all responses within subsequent analyses. While some of these responses may reflect respondent's lack of comprehension of the different schemes, it is also reasonable to assume that some reflect a personal rationality that larger schemes may be less likely to proceed; a perception which has been linked with lower WTP (Powe and Bateman, 2004).

this factor so that comparison with the theory-driven model can illuminate the impact of income in this analysis. We then contrast the theory-driven model by a 'statistically-driven' model which supplements the former with ad-hoc variables regarding which theory has no generic expectation but which empirical regularities observed in the literature suggest should improve the fit to the data at survey sites. In the next subsection we then transfer the various functions, calculate transfer errors and contrast these with those arising from simple mean value transfers.

Value functions were estimated by pooling data across the five countries. Both parametric and non-parametric analyses were conducted. As both identify common patterns in the data we focus upon the more readily interpretable parametric and report non-parametric results in the Appendix to this paper.

As out data contains both non-zero and valid zero bids, we have a WTP distribution which is censored. Given this, we specify a panel Tobit regression model which allows both this censoring and the fact that each respondent provides us with two WTP answers. The structural equation for such a random effects Tobit model is

$$y_{it}^* = X_{ikt}\beta_k + \mu_i + \varepsilon_{it} \quad (1)$$

where y^* is a latent variable observed for values greater than zero and censored otherwise, k indexes the number of independent variables included in the model, i is the individual index and t indexes our repeated responses. The random disturbances can be combined to form the composite error term of the model written as $w_{it} = \mu_i + \varepsilon_{it}$ which is assumed to be normally distributed. The latent variable y_{it}^* represents respondent i's unobserved willingness to pay to improve water quality at choice t, whereas the observable censored dependent variable y_{it} assumes value y_{it}^* when $y_{it}^* > 0$ and zero when $y_{it}^* < 0$.

The panel specification of Equation (1) captures both inter- and intra-respondent variation in WTP as well as incorporating the effect of observable and unobservable variables. In the random effects model the unobservable or un-measurable factors that differentiate respondents are assumed to be characterized as randomly distributed variables. Observable variables are incorporated in the usual way. Therefore, the random effects model can be thought of as a regression model with a random constant term. We employ simulated maximum likelihood estimation procedures to obtained unbiased, consistent and efficient estimates of the parameters β_k[13]. Table 2 reports the resulting models[14].

Table 2: Results from different models specification using random effects Tobit panel data model.

	All theory driven variables except income	Full theory-driven model	Statistically-driven model (including ad-hoc empirical regularities)
Variable	Coefficient (s.e.)	Coefficient (s.e.)	Coefficient (s.e.)

[13] This model was estimated using the STATA10 package. An alternative Heckman model within an initial hurdle identifying non-protest from protest respondents was also estimated but failed to yield significant improvements over the model reported here. Further, in order to test the stability of benefit transfer results we run a weighted panel Tobit model that takes into account the difference in sub-sample sizes. Results were relatively similar to those presented here which are preferred for their parsimony and ease of interpretation.

[14] Results for further specifications are reported in the Appendix to this paper.

Constant	12.58 (3.995)	-1.95 (5.076)	5.45 (7.934)
Large improvement	9.07 (1.055)	9.32 (1.073)	9.30 (1.073)
Small improvement 1st	5.47 (1.381)	5.15 (1.404)	5.09 (1.404)
Distance to the improvement site (km)	-0.19 (0.073)	-0.22 (0.074)	-0.19 (0.074)
Distance to nearest substitute site (km)	0.16 (0.037)	0.14 (0.037)	0.14 (0.038)
Income (net household income in € per year)	-	0.0008 (0.0002)	0.0008 (0.0002)
Age of respondent (in years)	-	-	-0.32 (0.099)
Urban (respondent lives in urban area=1; otherwise =0)	-	-	9.93 (3.232)
Norway	13.37 (4.294)	15.24 (4.359)	17.02 (4.518)
UK	-14.88 (5.228)	-17.36 (5.254)	-16.47 (5.254)
Belgium	27.85 (4.447)	22.76 (4.773)	24.87 (4.938)
Lithuania	-64.63 (5.053)	-50.43 (5.537)	-49.86 (5.569)
Sigma μ	73.47	75.20	72.79
Sigma ε	23.39	23.22	23.21
Rho	0.91	0.91	0.91
No. of observations	5790	5474	5466
Number of censored observations	1593	1457	1455
Log-Likelihood	-23186	-22142	-22098
Wald chi^2(*K*=restriction for overall significance)	573 (8)	525 (9)	552 (11)
p-value	0.000	0.000	0.000

Notes:
The country dummy for Denmark is omitted making this the baseline from which country departures are to be interpreted
The good/order dummy Small improvement 2nd is omitted making this the baseline from which the Large improvement 1st, Large improvement 2nd and Small improvement 1st departures are to be interpreted.
Rho = var_μ/(var_μ+var_ε) and represents the percent contribution to total variance of the panel-level variance component.

The p-value of the models in Table 2 show that they are globally highly significant. The parameter estimates on all variables for which we have prior expectations are statistically significant (at α = 5% with most significant at α=1%) and conform to those theory derived priors. The move from the model excluding income to the full theory-driven model has relatively little impact on the variable parameters although it does change the intercept and, as we discuss subsequently, significantly reduces transfer errors.

Briefly reviewing the relationships reported in Table 2, both scope sensitivity to changes in the quantity of improvement provided and any ordering effect are inspected by assigning the Small improvement 2nd responses as our base case WTP values. As can be seen the values accorded to the Large improvement are very clearly larger than this base case suggesting clear scope sensitivity. Both parametric and non-parametric ordering tests[15] confirmed there was no significant difference in this effect between the Large improvement 1st and Large improvement 2nd values and so these have been pooled. However, values for the Small improvement 1st, although below those for the Large improvement, are significantly above those for the Small improvement 2nd base case. This confirms the presence of the procedural invariance suspected from our inspection of Table 1. Very few benefit transfer studies conduct such tests and this finding suggests that such analyses may be worthwhile[16]. Given this result, in our subsequent function transfers we not only allow for the scope sensitivity difference between the Small and Large improvement but allow for ordering effects by conduct separate analyses for both the Small improvement 1st and Small improvement 2nd values.

Continuing with our inspection of Table 2, we find significant distance decay; as the distance between the respondent's home and the improvement site increases so WTP decays. Expectations are also borne out with respect to the substitution effect with WTP for the improvement site increasing as the distance to the nearest substitute rises. Income also has the expected positive impact on WTP.

The statistically-driven model extends the former analysis by adding in two ad-hoc variables for which economic theory has no prior expectations but which have been incorporated within previous analyses. Both the respondent's age and whether they live in an urban area are found to be significant predictors of WTP and result in an increase in the degree to which this model fits the data. From a statistical perspective these are stronger models of the study site data, however our contention is that their inclusion of such ad-hoc variables may increase transfer error relative to the full theory-driven model which only contains generic variables.

The *sigmas* represent the variances of the two error terms $\mu_{i\,i}$ and ε_{it}. Their relationship is described by the variable *rho*, which informs us about the relevance of the panel data nature. If this variable is zero, the panel-level variance component is irrelevant, but as can be seen from the results in Table 2, the panel data structure of the WTP answers has to be taken into account and is retained for all subsequent value function transfers to which we now turn.

4.3 Value transfer and error analyses

We conduct both simple mean value transfers and value function transfers for both our full dataset[17], including the dissimilar country (Lithuania) and excluding this to focus solely upon

[15] Non-parametric findings are reported the Appendix to this paper.

[16] It is worthwhile briefly considering why this may have arisen. One possibility is that this reflects a partial failure of the 'exclusive list' format resulting in a perceived change in the incentive compatibility of questions between the first and second response. It may also reflect the arguments of Carson and Groves (2007) regarding incentive properties of repeated valuation questions.

[17] We also conducted a number of preliminary transfers of non-pooled, individual country, values for each scope and ordering of the good. Non-parametric tests of the Kristofersson and Navrud (2004) equivalence hypothesis (whether we can take the mean WTP value of a randomly chosen country and validly transfer that to any other country in the sample) are generally rejected at a 20% error rate (ibid.). Similarly, simple mean transfers (taking one country as the policy site and another as the study site) are generally rejected. The only exceptions are the values of the large and small improvement in Denmark and Norway and the large improvement value in Denmark and Belgium. These results do not change if we account for the question ordering. Further details are given in the Appendix to this paper.

the remaining more similar countries. Value function transfers are undertaken using both theory-driven and statically-driving models. Together this allows us to test all of our various hypotheses regarding the appropriate methodology for value transfers in different contexts.

Mean value transfers are relatively straightforward being undertaken by pooling data from all countries except that which we are transferring to; the former being our 'study' sites and the latter the 'policy' site. Transfer errors are then calculated as an absolute percentage by comparing the mean value from the study sites with the actual mean of the policy site. We repeat this for each scope/ordering combination.

Value function transfers again compare the observed value of the 'policy' site[18] with that predicted from the other 'study' sites. However, now this prediction is obtained for each country in turn by estimating a value function (such as those shown in Table 2) on data from the study sites (i.e. omitting data from the policy site), then applying the coefficients[19] to the values of the predictors at the policy site to yield a function transfer value for that policy site[20]. Defining this value as $\hat{WTP}_{k|s}$ for policy site *k* estimated from study sites *s* and the directly observed policy site mean value as $\overline{WTP}_k$ then we calculate the value function transfer error as $\left(\frac{\hat{WTP}_{k|s} - \overline{WTP}_k}{\overline{WTP}_k}\right)\%$.

Table 3 presents the transfer errors obtained from both mean and value function transfers when we consider our full dataset including the dissimilar (Lithuanian) site. The upper section of this table details results for the mean value transfers, disaggregated into values for the Large and Small improvement with the latter further disaggregated to allow for the significant framing (ordering) effect observed in these values. We can see that this mix of similar and dissimilar sites yields high levels of overall error (shown in the final column) when the mean value approach is applied with an overall raw error rate of 116%. This is likely to be unacceptably high for policy purposes and so function transfers seem worthy of investigation. However, as an aside, even a cursory inspection of the mean value transfer results for individual countries clearly bears out our expectation that it is the dissimilar Lithuanian site which is the principle cause of these high error rates. Subsequently we investigate the impact of restricting our analysis to just the similar countries.

---

[18] Some analysts compare value function estimates from study sites with the mean value from the policy site (e.g. Van den berg et al.2001, Brouwer and Bateman 2005). However, others compare the former value with that predicted by a function estimated from the policy site data (e.g. Barton 2002; Chattopadhyay 2003). In the present study we tested both approaches but found no difference in the pattern of results provided. Consequently we report the mean value comparisons here (directly comparable with the simple univariate transfers) and present the comparisons with predicted value in the Appendix to this paper.

[19] Given the Tobit specification, the parameters to be used for function transfer must be adjusted for censoring. Discussions of this adjustment can be found in Halstead et al. (1991), Haab and McConnell (2003) and Brouwer and Bateman (2005).

[20] An interesting issue is whether one should apply the estimated coefficients to the mean value of the policy site predictors or to the values that apply for each individual. Normally the former approach is the only one available and we follow that method here. However, all of the predictors in our theory-driven model can be assessed at the individual level using secondary source data (from GIS and census datasets). This raises the possibility that one could synthesise a dataset for an unsurveyed policy site consisting of the change in provision, distance to the improvement site, distance to substitutes and household income. One could then apply coefficients from a value function transfer at this individual response level thus obviating the need to rely upon mean predictor values in the transfer process. We suspect that this may further reduce value function transfer errors.

Table 3: Transfer errors (%) from mean value and function transfer methods: Dataset including the dissimilar (Lithuanian) site.

WTP measure[1]	Lithuania	Belgium	Denmark	Norway	UK	Average errors (weighted)[2]
MEAN VALUE TRANSFERS						
Small improvement 1st	508	49	6	41	48	130 (102)
Small improvement 2nd	392	53	23	43	69	116 (94)
Large improvement	391	39	10	37	34	102 (81)
Average error (weighted)	430 (420)	47 (45)	13 (12)	40 (40)	50 (46)	**116** (90)
VALUE FUNCTION TRANSFERS						
Reduced theory-driven model (Provision change, distance to site and distance to substitute)						
Small improvement 1st	69	100	98	95	89	90(84)
Small improvement 2nd	125	106	116	101	104	111(101)
Large improvement	48	94	96	91	84	82 (78)
Average error (weighted)	80 (72)	100 (98)	103 (102)	96 (94)	92 (90)	**94** (88)
Full theory-driven model (Provision change, distance to site, distance to substitute and income)						
Small improvement 1st	30	74	65	80	43	58(62)
Small improvement 2nd	81	76	71	83	40	70(71)
Large improvement	13	66	69	75	43	53 (58)
Average error (weighted)	41 (34)	72 (71)	68 (68)	79 (78)	42 (42)	**61** (64)
Statistically-driven model (Provision change, distance to site, distance to substitute, income, age and urban)						
Small improvement 1st	35	81	93	101	72	77 (81)
Small improvement 2nd	78	84	103	107	78	90 (92)
Large improvement	12	74	86	94	64	66(72)
Average error (weighted)	41(34)	80 (78)	94(92)	101(99)	71(70)	**77** (82)

Notes: Transfer errors calculated by comparison with the mean WTP values estimated at each site.

1. Small improvement 1st = WTP for the smaller improvement elicited as the first valuation question asked; Small improvement 2nd = WTP for the smaller improvement elicited as the second valuation question asked (significantly lower than Small improvement 1st); Large improvement 1st or 2nd= WTP for the larger improvement elicited as either the first or second valuation question asked (no significant difference between these responses).

2. Raw averages given outside parentheses. Figures in parentheses are average errors weighted by relative sample size

The remainder of Table 3 details results for our various function transfer analyses. This starts with the 'reduced theory-driven' model containing all of those variables suggested by economic theory except for the major source of dissimilarity between countries; income. As can be seen, even this model generates a substantial reduction in overall error relative to the mean value approach. This error reduction is further improved when we add in the income variable to specify our full theory-driven model. Given that this is the major source of dissimilarity across sites it is not surprising that this variable generates a larger improvement than any of the others in this model which nearly halves the rate of overall error generated by the mean value approach. We now test our hypothesis that the theory driven model, although not providing quite such a good fit to study site data as the statistically driven model (recall the results Table 2) will nonetheless provide a lower rate of transfer error than the latter. To test this we now calculate transfer errors for the statistically driven model. These are more than one quarter higher than the transfer errors associated with the theory driven model. The only change here is the addition of the ad-hoc variables, not prescribed by theory as being generic components of utility functions. This, we contend, provides strong support for the methodological principles proposed at the start of this paper.

The results presented in Table 3 support the hypothesis that, when faced with a heterogeneous set of sites, analysts should prefer function transfer over mean value transfer (and within the former should restrict the specification of models to those variables regarding which economic theory has clear expectations). However, the individual country results show that mean value transfers can yield both high rates of error when transfers from similar countries are applied to a dissimilar country (Lithuania) but lower rates when (mainly) similar countries are used to predict for other similar countries. To investigate the potential for error reduction here we now exclude the dissimilar Lithuanian case study and repeat the previous analyses for the remaining similar countries, results being reported in Table 4.

Table 4: Transfer errors (%) from mean value and function transfer methods: Dataset excluding the dissimilar (Lithuanian) site.

WTP measure[1]	Belgium	Denmark	Norway	UK	Average errors (weighted)[2]
MEAN VALUE TRANSFERS					
Small improvement 1st	30	45	22	91	47 (40)
Small improvement 2nd	35	67	24	119	61 (51)
Large improvement	19	19	19	69	32 (27)
Average error (weighted)	28 (26)	44 (38)	22 (21)	93 (87)	**47** (40)
VALUE FUNCTION TRANSFERS					
Reduced theory-driven model (Provision change, distance to site and distance to substitute)					
Small improvement 1st	92	85	89	82	87 (88)
Small improvement 2nd	97	101	94	96	97 (97)

Large improvement	84	86	83	77	83 (83)
Average error (weighted)	91(90)	91(89)	89(87)	85(83)	**89** (89)
Full theory-driven model (Provision change, distance to site, distance to substitute and income)					
Small improvement 1st	79	65	81	53	69 (72)
Small improvement 2nd	82	72	84	54	73 (76)
Large improvement	70	68	75	50	66 (68)
Average error (weighted)	77(75)	69 (69)	80 (79)	52 (52)	**69** (72)
Statistically-driven model (Provision change, distance to site, distance to substitute, income, age and urban)					
Small improvement 1st	74	67	92	47	70 (75)
Small improvement 2nd	76	67	97	43	71 (76)
Large improvement	65	63	84	41	63 (68)
Average error (weighted)	72 (70)	66(65)	91(90)	44(43)	**68** (73)

Notes: Transfer errors calculated by comparison with the mean WTP values estimated at each site.

1. Small improvement 1st = WTP for the smaller improvement elicited as the first valuation question asked; Small improvement 2nd = WTP for the smaller improvement elicited as the second valuation question asked (significantly lower than Small improvement 1st); Large improvement 1st or 2nd= WTP for the larger improvement elicited as either the first or second valuation question asked (no significant difference between these responses).
2. Raw averages given outside parentheses. Figures in parentheses are average errors weighted by relative sample size.

Comparison of Tables 3 and 4 clearly shows that the exclusion of the dissimilar Lithuanian sites dramatically reduces the rate of error associated with mean value transfers such that they now fall well below those achieved by value function transfers. Including the dissimilar data reverses this relationship. This supports our central hypothesis that the choice of method depends crucially upon the degree of similarity of the sites under consideration; an issue which we now discuss further in our concluding remarks.

5. Conclusions

The central objective of this study was to develop principles for choosing between the different methodologies available for conducting benefit transfers. Our analysis shows that the crucial issue concerns the degree of heterogeneity between the various sites across which transfers are to be undertaken. Results show that the pertinent dimensions of similarity or difference can be assessed using data (which is readily available from secondary sources) regarding the characteristics of those sites and their surrounding populations. Using such we find that when analysis is restricted to only include similar sites transfer errors are minimised when simple mean value methods are applied. However, when we also included dissimilar sites then mean value approaches generated the highest levels of error. In such cases function transfer approaches addressed the higher heterogeneity of such a group of sites and provided lower errors. These errors were minimised when we specified transfer functions to only include those generic variables which economic theory expects to be present in typical utility functions. Specifying value functions to include ad-hoc variables for which theory has no expectations resulted in an improvement in the statistical fit of those functions to study site data but led to higher error when those functions were transferred to predict value at other sites.

The results presented in this paper provide straightforward principles for the application of benefit transfer. When transferring across similar sites, mean value approaches are to be preferred. When transferring across heterogeneous sites, value function transfers are preferable and the specification of those functions should be restricted to include only those generic variables for which we have prior economic expectations. Such principles, we argue, should provide guidelines for future applications in this expanding field.

References

Ariely, D., Loewenstein, G. and Prelec, D. (2003). 'Coherent arbitrariness': stable demand curves without stable preferences, *Quarterly Journal of Economics* 118 (1): 73–105.

Arrow, K., Solow, R., Portney, P.R., Leamer, E.E., Radner, R. and Schuman H. (1993). Report of the NOAA panel on contingent valuation, *Fed. Reg*. 58 (1993): 4601–4614.

Banerjee, S. and Murphy, J.H. (2005). The scope test revisited, *Applied Economics Letters* 12(10): 613-617.

Barton, D.N. (2002). The transferability of benefit transfer: contingent valuation of water quality improvements in Costa Rica. *Ecological Economics*, 42: 147-164.

Bateman, I.J. and Brouwer, R. (2006). Consistency and construction in stated WTP for health risk reductions. A novel scope sensitivity test. *Resource and Energy Economics* 28: 199-214.

Bateman, I.J., Burgess, D., Hutchinson, W.G. and Matthews, D.I. (2008). Contrasting NOAA guidelines with Learning Design Contingent Valuation (LDCV): Preference learning versus coherent arbitrariness. *Journal of Environmental Economics and Management* 55: 127–141. http://dx.doi.org/10.1016/j.jeem.2007.08.003

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

Bateman, I.J., Cole, M., Cooper, P., Georgiou, S., Hadley, D. and Poe, G.L. (2004). On visible choice sets and scope sensitivity, *Journal of Environmental Economics and Management* 47: 71-93. DOI: 10.1016/S0095-0696(03)00057-3.

Bateman, I.J., Day, B.H., Dupont, D. and Georgiou, S. (forthcoming). Procedural invariance testing of the one-and-one-half-bound dichotomous choice elicitation method. *Review of Economics and Statistics*, in press.

Bateman, I.J., Day, B.H., Georgiou, S. and Lake, I. (2006). The aggregation of environmental benefit values: Welfare measures, distance decay and total WTP. *Ecological Economics*, 60(2): 450-460, doi:10.1016/j.ecolecon.2006.04.003.

Bateman, I.J., Day, B.H., Jones, A. P. and Jude, S. (2009) Reducing gains/loss asymmetry: A virtual reality choice experiment (VRCE) valuing land use change, *Journal of Environmental Economics and Management*, 58: 106-118, doi:10.1016/j.jeem.2008.05.003

Bateman, I.J., Dent, S., Peters, E., Slovic, P. and Starmer, C. (2007). The Affect Heuristic and the Attractiveness of Simple Gambles, *Journal of Behavioral Decision Making* 20(4): 365-380. DOI: 10.1002/bdm.558.

Bateman, I.J. and Jones, A.P. (2003). Contrasting conventional with multi-level modelling approaches to meta-analysis: An illustration using UK woodland recreation values, *Land Economics* 79(2): 235-258. DOI:10.3368/le.79.2.235.

Bateman, I.J., Jones, A.P., Lovett, A.A., Lake, I. and Day B.H. (2002b) Applying geographical information systems (GIS) to environmental and resource economics, *Environmental and Resource Economics*, 22(1-2): 219-269. **DOI:** 10.1023/A:1015575214292

Bateman, I.J., Langford, I.H., Turner, R.K., Willis, K.G. and Garrod, G.D. (1995). Elicitation and truncation effects in contingent valuation studies, *Ecological Economics* 12(2):161-179. DOI: 10.1016/0921-8009(94)00044-V.

Bateman, I.J., Kahneman, D., Munro, A., Starmer, C. and Sugden, R. (2005). Testing competing models of loss aversion: An adversarial collaboration, *Journal of Public Economics* 89(8): 1561–1580, doi:10.1016/j.jpubeco.2004.06.013.

Bateman, I.J., Munro, A., Rhodes, B., Starmer, C. and Sugden, R. (1997a). A test of the theory of reference-dependent preferences, *Quarterly Journal of Economics* 112(2): 479-505.

Bateman, I.J., Munro, A., Rhodes, B., Starmer, C. and Sugden, R. (1997b). Does part-whole bias exist? An experimental investigation, *Economic Journal* 107(441): 322-332.

Bergland, O., Magnussen, K. and Navrud, S. (1995). Benefit transfer: testing for accuracy and reliability. *Discussion Paper, #D-03/1995*, Department of Economics and Social Sciences. Agricultural University of Norway.

Birnbaum, M. H. (1999). How to show that 9 > 221: Collect judgments in a between-subjects design. *Psychological Methods* 4(3): 243–249.

Brouwer, R. (2000). Environmental Value Transfer: State of the Art and Future Prospects. *Ecological Economics* 32: 137-152.

Brouwer, R. and Bateman, I.J. (2005). Benefits Transfer of Willingness to Pay Estimates and Functions for Health-Risk Reductions: A Cross-Country Study. *Journal of Health Economics* 24: 591-611.

Brouwer, R. and Spaninks, F.A. (1999). The validity of environmental benefits transfer: Further empirical testing. Environmental and Resource Economics 14(1): 95-117.

Camerer, C. F., Loewenstein, G. & Rabin, R. (eds.) (2003) *Advances in Behavioral Economics*.

Carson, R.T., Flores, N.E. and Hanemann, W.M. (1998) Sequencing and valuing public goods, *Journal of Environmental Economics and Management* 36: 314-323.

Carson, R.T. and Groves, T. (2007). Incentive and informational properties of preference questions, *Environmental & Resource Economics* 37(1): 181-210.

Carson, R., Hanemann, M., Kopp, R., Krosnick, J., Mitchell, R., Presser, S., Rudd, P. and Smith, V.K. with Conaway, M. and Martin, K. (1996) Was the NOAA panel correct about contingent valuation?, *Resources for the Future Discussion Paper 96-20*, Washington D.C.

Carson, R. T. and Mitchell, R. C. (1993). The value of clean water: The public's willingness to pay for boatable, fishable and swimmable quality water, *Water Resources Research*, 29(7): 2445–2454.

Champ, Patty A., Boyle, Kevin and Brown, Tom C. (eds.) (2003). *A Primer on Non-market Valuation*, The Economics of Non-Market Goods and Services: Volume 3, Kluwer Academic Press, Dordrecht.

Chattopadhyay S. (2003). A Repeated Sampling Technique in Assessing the Validity of Benefit Transfer in Valuing Non-Market Goods, *Land Economics*, 79 (4): 576–596.

Davies, H. and Neal, C. (2004). GIS based methodologies for assessing nitrate, nitrite and ammonium distributions across a major UK basin, the Humber, *Hydrology and Earth System Science* 8:823-833.

Covey, J., Loomes, G. and Bateman, I.J., (2007) Valuing risk reductions: Testing for range biases in payment card and random card sorting methods, *Journal of Environmental Planning and Management*, 50(4): 467-482. **DOI:** 10.1080/09640560701401986

DeShazo, J.R. (2002). Designing transactions without framing effects in iterative question formats, *Journal of Environmental Economics and Management* 43: 360–385.

Desvousges, W.H., Naughton, M.C. and Parsons, G.R. (1992). Benefit transfer: conceptual problems in estimating water quality benefits using existing studies. *Water Resources Research*, 28(3): 675-683.

Desvousges, W.H., Smith, V.K. and Fisher, A. (1987). Option price estimates for water quality improvements: a contingent valuation study of the Monongahela River*, Journal of Environmental Economics and Management* 14: 248-267.

Douglas, B.B. and Rangel, A. (2008). Behavioural public economics, *The New Palgrave Dictionary of Economics*, 2nd Edition. Macmillan.

Downing, M. and Ozuna, T. (1996). Testing the reliability of the benefit function transfer approach. *Journal of Environmental Economics and Management* 30: 316-322.

Doyle, J. R., O'Connor, D. J., Reynolds, G. M., & Bottomley, P. A. (1999). The robustness of the asymmetrically dominated effect: Buying frames, phantom alternatives, and in-store purchases, *Psychology & Marketing* 16(3): 225-243.

European Parliament (2000) Directive 2000/60/EC (The EU Water Framework Directive), *Official Journal* (OJ L 327): 22 December 2000.

Fagerlin, A., Wang, C. and Ubel, P. A. (2005). Reducing the Influence of Anecdotal Reasoning on People's Health Care Decisions: Is a Picture Worth a Thousand Statistics? *Medical Decision Making* 25:398–405.

Haab, T,C„ and K.E. McConnell. 2003. *Valuing Environmental and Natural Resources: The Econometrics of Nonmarket Valuation*. Edward Elgar Publishing, Massachusetts.

Halstead J.M., Lindsay B.E., Brown C.M., 1991. Use of the Tobit model in the contingent valuation: experimental evidence from the Pemigewasset wilderness area. *Journal of Environmental Management* 33: 79:89.

Hime, S., Bateman, I.J., Posen, P. and Hutchins, M. (2009) A transferable water quality ladder for conveying use and ecological information within public surveys, *CSERGE Working Paper EDM 09-01,* Centre for Social and Economic Research on the Global Environment, University of East Anglia.

Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes* 67: 247–257.

Hutchins, M. G., Deflandre, A. and Boorman, D. B. (2006). Performance benchmarking linked diffuse pollution and in-stream water quality models. *Archiv fur Hydrobiologie Supplement*, 161 (Large Rivers Supplement 17): 133-154.

Johnston, R.J., Besedin, E.Y., Iovanna, R., Miller, C.J., Wardwell, R.F. and Ranson, M.H. (2005). Systematic variation in willingness to pay for aquatic resource improvements and implications for benefit transfer: A meta-analysis, *Canadian Journal of Agricultural Economics* 53: 221-48.

Johnston, R.J., Ranson, M.H., Besedin, E.Y. and Helm, E.C. (2006). What determines willingness to pay per fish? A meta-analysis of recreational fishing values, *Marine Resource Economics* 21: 1-32.

Jones, A.P., Bateman, I.J. and Wright, J. (2002). *Estimating arrival numbers and values for informal recreational use of British woodlands*, report to the Forestry Commission, Edinburgh, published at http://www.forestry.gov.uk.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist* 58: 697–720.

Kahneman, D. and Knetsch, J.L. (1992). Valuing public goods: The purchase of moral satisfaction, *Journal of Environmental Economics and Management* 22(1): 57-70.

Kahneman, D., Slovic, P. and Tversky, A. (1982). *Judgement Under Uncertainty: Heuristics and Biases*, Cambridge University Press, New York.

Leon-Gonzalez, R. and Scarpa, R. (2008). Improving multi-site benefit functions via Bayesian model averaging: A new approach to benefit transfer**,** *Journal of Environmental Economics and Management* 56(1): 50-68.

Lichtenstein, S., and Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology* 89: 46–55.
Lindhjem, H.L. and Navrud, S. (2008). How reliable are meta-analyses for international benefit transfers? *Ecological Economics* 66: 425-435.
List, J.A. (2002). Preference Reversals of a Different Kind: The "More Is Less" Phenomenon, *American Economic Review* 92(5): 1636-1643.
Mitchell, R.C. and Carson, R.T. (1989). *Using Surveys to Value Public Goods: the Contingent Valuation Method*, Resources for the Future, Washington D.C.
Moeltner, K., Boyle, K.J. and Paterson, R.W. (2007). Meta-analysis and benefit transfer for resource valuation–Addressing classical challenges with Bayesian modeling, *Journal of Environmental Economics and Management* 53: 250-69.
Mullainathan, S. and Thaler, R. H. (2001). Behavioral Economics, *International Encyclopedia of the Social & Behavioral Sciences*. pp. 1094–1100. doi:10.1016/B0-08-043076-7/02247-6.
Muthke, T. and Holm-Mueller, K. (2004). National and international benefit transfer testing with a rigorous test procedure. *Environmental and Resource Economics* 29: 323-336.
Navrud, S. and Ready, R. (2007a). Review of methods for value transfer. In: Navrud, S. and Ready, R. (eds.), *Environmental value transfer: Issues and methods*. Springer, Dordrecht, The Netherlands.
Navrud, S. and Ready, R. (eds.) (2007b). *Environmental value transfer: Issues and methods*. Springer, Dordrecht, The Netherlands.
Pearce, D.W., Whittington, D. and Georgiou, S. (1994). *Project and policy appraisal: Integrating economics and environment*. OECD, Paris.
Peters, E., Slovic, P. and Hibbard, J. (2005). *Bringing Meaning to Numbers: Evaluability and Affect in Choice*, University of Oregon, Eugene, Oregon.
Powe, N.A. and Bateman, I.J. (2004) Investigating insensitivity to scope: A split-sample test of perceived scheme realism, *Land Economics*, 80(2): 258-271.
Rabin, M. (1998). Psychology and Economics. *Journal of Economic Literature* 36 (1): 11–46. doi:10.2307/2564950.
Ready, R., Navrud, S., Day, B., Dubourg, R., Machado, F., Mourato, S., Spaninks, F. and Vázquez Rodriquez, M.X. (2004). Benefit transfer in Europe: How reliable are transfers between countries? *Environmental and Resource Economics*, 29: 67-82.
Rowe, R.D., Schulze, W.D. and Breffle, W. (1996) A test for payment card biases, *Journal of Environmental Economics and Management*, 31: 178-185.
Santos, J.M.L. (2007). Transferring landscape values: how and how accurately? in Navrud, S. and Ready, R. (Eds.) *Environmental Value Transfer: Issues and Methods*, The Economics of Non-Market Goods and Services: Volume 9, Springer, Dordrecht.
Scarpa, R., Hutchinson, W.G., Chilton, S.M. and Buongiorno, J. (2007) Benefit value transfers conditional on site attributes: Some evidence of reliability from forest recreation in Ireland, in Navrud, S. and Ready, R. (Eds.) *Environmental Value Transfer: Issues and Methods*, The Economics of Non-Market Goods and Services: Volume 9, Springer, Dordrecht.
Schkade, D. A. and Payne, J. W. (1994). How people respond to contingent valuation questions: A verbal protocol analysis of willingness to pay for an environmental regulation. *Journal of Environmental Economics and Management* 26: 88–109.
Simon, H. A. (1987). Behavioural economics, *The New Palgrave: A Dictionary of Economics*, v. 1, pp. 221–24.
Slovic, P. (1995). The construction of preference. *American Psychologist* 50: 364–371.
Tversky, A. and Kahneman, D. (1974). Judgement under uncertainty: heuristics and biases, *Science* 185: 1124–1131.
UKTAG (2008). UK Environmental Standards and Conditions (Phase 1): Final Report - April 2008, The UK Technical Advisory Group on the Water Framework Directive (UKTAG), http://www.wfduk.org/
Van den Berg TP, Poe GL, and Powell JR. 2001. Assessing the accuracy of benefits transfers: evidence from a multi-site contingent valuation study of ground water quality.

In: Bergstrom J, Boyle K, and Poe G (Eds). *The Economic Value of Water Quality*. Northampton, MA: Edward Elgar.

WATECO (2004) *Economics and the Environment - The Implementation Challenge of the Water Framework Directive.* Luxembourg: European Commission.

World Bank (2008). Global purchasing power parities and real expenditures. World Bank, Washington, D.C., U.S.A.

Zikmund-Fisher, B J., Fagerlin, A. and Ubel, P. A. (2005). What's Time Got to Do with It? Inattention to Duration in Interpretation of Survival Graphs, *Risk Analysis* 25(3): 589-595.

Zandersen M., Termansen M., Jensen F.S. (2007) Testing Benefits Transfer of Forest Recreation Values over a Twenty-Year Time Horizon, *Land Economics*, 83 (3): 412–440.

Appendices

This section provides the various additional materials referred to in the main text of this paper.

Appendix. Common design principles for value transfers

Good practice within individual case studies, while necessary, may not be sufficient within the value transfer context. One of the basic requirements which has been stressed from the early days of value transfer, is the need for a common design format for the source studies used in developmental work (Desvousges et al., 1992). We can identify a series of principles for designing studies for subsequent value transfer and their subsequent analysis, as follows:

- Study design should be developed from economic theoretic principles and employ a theoretically consistent utility specification;
- The design should permit robust validity testing[21]
- Survey questionnaires must be appropriate and common to all study and foreseeable policy sites with questions being consistently phrased with common response options and levels and common data coding;
- Source valuation studies should ideally elicit information on the location of the good and the location of respondents so that distance decay relationships can be assessed. Substitute availability should either be assessed via the source studies or during the transfer exercise.
- Information on the do-nothing and post provision change scenarios must be readily comprehended, adequate and common to all study site applications;
- The valuation method should either be common to all studies or vary sufficiently to ensure that methodological effects can be assessed[22];
- The framing of valuation questions and the method through which valuation responses are elicited should either be common or should vary sufficiently to ensure that the effects of such variation can be assessed[23];
- Mean value transfer tests should be carried out. Results reported in the main body of this paper suggest that these will minimise error where the study and policy sites are similar in terms of the good provided, the characteristics of the site and the characteristics of the relevant population.
- Where transfers include dissimilar sites then value function transfer methods may well outperform simple mean value transfers.
- Empirical specifications of value functions should be theory driven, being developed from the specified structure of a theoretically consistent utility model rather than including ad-hoc, context specific variables chosen primarily for their contribution to statistical fit at any given site;
- Similarly analysts should avoid the lure of over-reliance upon readily available, context specific, survey derived variables such as those concerning the minutiae of

[21] This testing should include both transfer error analyses and tests based on theoretically derived expectations. As discussed subsequently we consider the common 'scope sensitivity' test (Arrow et al., 1993) to be necessary but insufficient for this purpose.

[22] What should be avoided is that studies used for transfer purposes apply a small number of methods with just a few studies for each method. In such cases methodological and study site effects are difficult to separate.

[23] In theory once a 'correct' framing and elicitation method are identified then there is no need to test for the effects of varying these design features. However, as the literature regarding elicitation options in the CV method illustrates, consensus can be elusive and debate prolonged (e.g. Bateman et al., 1995; Carson and Groves, 2007).

respondent characteristics and instead focus inclusion of those theoretically derived variables (e.g. income, use, substitute availability) which are liable to be common across sites;

- GIS analyses provide ready quantification of off-site locational issues such as distance from respondent's home to site (proxying use in a readily transferable manner and allowing the estimation of distance decay in values[24]) and the availability of substitutes (again via distance measures). GIS also facilitates the ready transfer of functions containing such variables to other policy sites;
- Where transfers are conducted across countries (or even over large distances if relevant and data are available) then appropriate purchasing power parity (PPP) adjustments should be undertaken prior to common analyses of data.

[24] As discussed in Bateman et al.,(2006), quantification of distance decay is vital to the accurate aggregation of WTP values, avoiding the overestimates which arise when mean values derived from samples which are often collected near to sites are applied to wider populations. Indeed distance decay functions also allow the analyst to define the 'economic jurisdiction'; that spatial area beyond which the value of improvements falls to zero. Again explicit recognition of this area avoids overestimation of total WTP values.

Appendix: Developing the water quality ladder

A vital element of any stated preference valuation study is the definition of the good concerned, its status quo conditions and the change in provision which will we will ask survey respondents to value. This in turn requires an understanding of the physical science determining these states. While there are numerous pollutants that affect open access waters, the WFD focuses upon those which affect their ecological status and in particular those nutrients that are delivered to waterways via routes such as diffuse pollution from agriculture (Davies and Neal, 2004; Hutchins et al., 2006; Neal and Jarvie, 2005; Neal et al., 2005)[25]. However, individuals do not hold values for reducing pollution per se but rather for the effects that such reductions may induce in terms of recreation suitability and ecological quality. To some considerable degree the pathways linking pollution to ecological impact is still the subject of ongoing research (UKTAG, 2008). However, this does not prevent the analyst from valuing certain states of the world on the assumption that ongoing research will indicate how such states might subsequently be attained. With this in mind Hime and Bateman (2009) define four levels of river water quality based upon chemical, physical, flora and fauna characteristics. Table A1 provides details of this characterisation exercise.

Table A1: Chemical, physical, flora and fauna based characterization of river water quality.

Highest quality **BLUE**	**GREEN**	**YELLOW**	**Lowest quality** **RED**
Chemistry			
BOD Limit < 4mgl^{-1} Cat: A & B	BOD Limit >= 4mgl^{-1} and < 6mgl^{-1} Cat: C	BOD Limit >= 6 and < 8mgl^{-1} Cat: D	BOD > 8mgl^{-1} Cat: E & F
Freshwater fish directive limit game BOD Limit = 3 mgl^{-1}		Freshwater fish directive limit BOD Limit = 6	
Ammonia < 0.6 mgNl^{-1}	Ammonia < 1.3 mgNl^{-1}	Ammonia < 2.5 mgNl^{-1}	Ammonia > 2.5mgNl^{-1}
Assumed physical state			
Patches of faster flow	Lower flow rate; no fast patches	Low flow rate	Very low flow rate
Gravel / pebble substrate; No algae on rocks	Small gravel and sand substrate; little algae on rocks	Mud; algae on rocks	Mud; algae on rocks
Aquatic plants			
No algae; Water plants (described below); Good clarity	Greater amount of aquatic plants taking up more of the open space; Slight increase in water turbidity	Less aquatic plants with increases in algae; Further increase in turbidity and green hue to the water, Small number of algal mats	Large degree of siltation; Turbid water with a brown hue; Algal mat covering the substrate
Total vegetation cover = 50%	Total vegetation cover = 60%	Total vegetation cover = 70%	Total vegetation cover = 85%
Rhynchostegium	*Apium nodiflorum*	*Apium nodiflorum*	algae *Cladopora*

[25] In other research under the ChREAM program (Bateman et al., 2006) we seek to extend the work of Kay et al., (2005) in examining the link between land use change and water borne faecal matter.

riparoides (20); *Myriophyllum alterniflorum*[1] (20); *Leptodictyum (Amblystegium) fluviatile* (10); *Fontinalis antipyretica* (10) *Ranuculus penicillatus ssp. Pseduofluitans*[1] (4); *Pellia endiviifolia* (2); *Apium nodiflorum* (3); *Callitriche hamulata*[1] (10); *Leptodictyum (Amblystegium) riparium* (3); *Rorippa nasturtium-aquaticum* (3) *Callitriche platycarpa*[1] (5); *Callitriche stagnalis*[1] (2); *Potamogeton crispus* (2); *Potamogeton natans*[2] (6)	(20); *Leptodictyum (Amblystegium) riparium* (20); *Potamogeton crispus* (10) *Rhynchostegium riparoides* (15); *Myriophyllum alterniflorum*[1] (10); *Leptodictyum (Amblystegium) fluviatile* (5); *Fontinalis antipyretica* (5) *Callitriche hamulata*[1] (2); *Callitriche stagnalis*[1] (8); *Potamogeton crispus* (5)	(5); *Leptodictyum (Amblystegium) riparium* (50); *Potamogeton crispus* (5) Algae *Cladopora etc. (40)*	*etc.(100)*
Fish – general assessment			
Game and coarse	Same or higher coarse numbers, few game fish	Lower coarse fish, no game fish.	Very few fish
Fish – species breakdown			
Brown trout (mid) central area fastest flow	-	-	-
Minnow (high)	-	-	-
Vendace (mid)	-	-	-
Barbel (mid)	-	-	-
Chub (mid)	-	-	-
-	Bream	Bream	-
-	Common Carp (mid) mid-water	Common Carp (low) Whole area – not edges (silt)	-
-	Perch (less) mid-water	-	-
-	Roach (mid) mid-water	Roach (high) Whole area – not edges (silt)	-
-	Rudd (mid) mid-water	Rudd (low) Whole area – not edges (silt)	-

Pike (v. low)	Pike (v. low) mid-water	Pike (v. low) Whole area – not edges (silt)	-
-	-	Stickle Back (mid) edges as small fish, not where there is too much silt	-
Uses			
Game fishing	-	-	-
Coarse fishing	Coarse fishing	Restricted coarse fishing	-
Swimming	Swimming	-	-
Canoeing & boating	Canoeing & boating	Canoeing & boating	-
Bird watching	Bird watching	Bird watching	Restricted bird watching

Source: adapted from Hime and Bateman (2009)

Biological oxygen demand (BOD) and ammonia levels from UKTAG (2008) and EA (2007a,b). Aquatic plant frequency and species from Holmes et al., (1999) and JNCC (2005). 1 = Aquatic plant species which occur at up to 0.5m depth (EA, 2007c,d,e); 2 = Aquatic plant species which occur at 0.5 – 1.5m depth (EA, 2007c,d,e). Numbers in parentheses to the left of plant community composition show the percentage breakdown of the total vegetation cover. Physical assessments and fish species information from EA (2007f,g).

The complexity of information given in Table 1 is of course far too high to reasonably allow its unadjusted use as survey information. Therefore, following discussions across the various case study partners, a set of photographs of generic water quality characteristics was agreed for each quality level. These were then passed to a graphic artist to produce the generic water quality ladder[26] shown in the main text of this paper. Qualitative face-to-face testing with a pilot sample confirmed that this form of information was clearly comprehended by respondents who were able to recall patterns in quality change following the interview process.

[26] Note that this version of the water quality ladder ties together the ecological and use quality of rivers. This need not be the case as the drivers of ecological quality (mainly nutrients etc.) are, within reasonable limits, not those which determine suitability for use (E Coli levels and other faecal matter; see Kay et al., 2005). Furthermore, these various drivers need not be correlated (although in practise they frequently are). Therefore, in ongoing valuation work we break the deterministic link between ecological quality and use suitability shown in Figure 1. Nevertheless, our expectation is that, in terms of preferences, use and ecological utilities may well be empirically correlated (e.g. individuals dislike direct contact with water which has high algae levels even if they have low faecal matter and low health risk). Given this it may be that our estimated values are not contingent upon the ecology / use link specified in Figure 1.

Appendix: Valuation questions

All countries adopted a common questionnaire. The UK questionnaire was computerized with a touch screen helping to identify the location of river stretches and recreational visits. The valuation section from that questionnaire is reported verbatim below.

<u>Extract from UK survey questionnaire</u>

I now want to ask a different type of question. Please look at this map; you live here [POINT TO MAP HOUSE ON MAP].

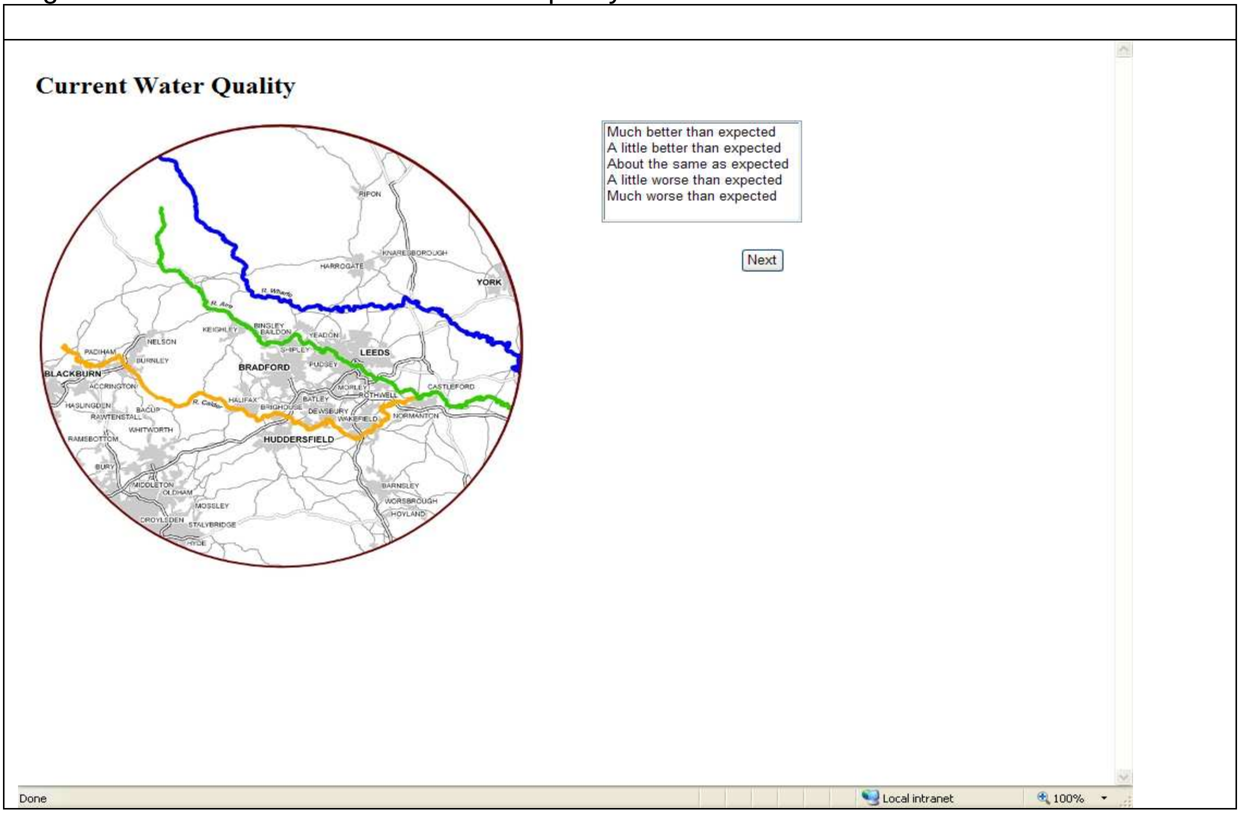
We have used the colours from the pictures to show the current water quality of rivers in this area. This is based on information from the Environment Agency, which is the official body that monitors river quality in the UK.

As you can see, at present the river closest to you [INDICATE RIVER CLOSEST TO RESPONDENTS HOME], is coloured [SAY COLOUR] which means that on average its water quality is like this [point to picture corresponding to colour]. This river [INDICATE 2ND CLOSEST RIVER TO RESPONDENTS HOME] is the next closest and is [STATE COLOUR OF THAT RIVER] on average its water quality is like this [POINT TO PICTURE CORRESPONDING TO COLOUR ON THE WATER QUALITY LADDER].

Finally, the furthest River [INDICATE THE RIVER 3RD CLOSEST TO RESPONDENTS HOME] is coloured [STATE COLOUR OF THAT RIVER]. So, on average, its water quality is like this [point to picture corresponding to colour].

E1. Looking at these categories [*SHOWCARD RIVERS QUALITY*], which phrase best, describes your reaction to the information concerning the general current water quality of rivers in the area?

Figure A.1: Screen shot - Current water quality levels



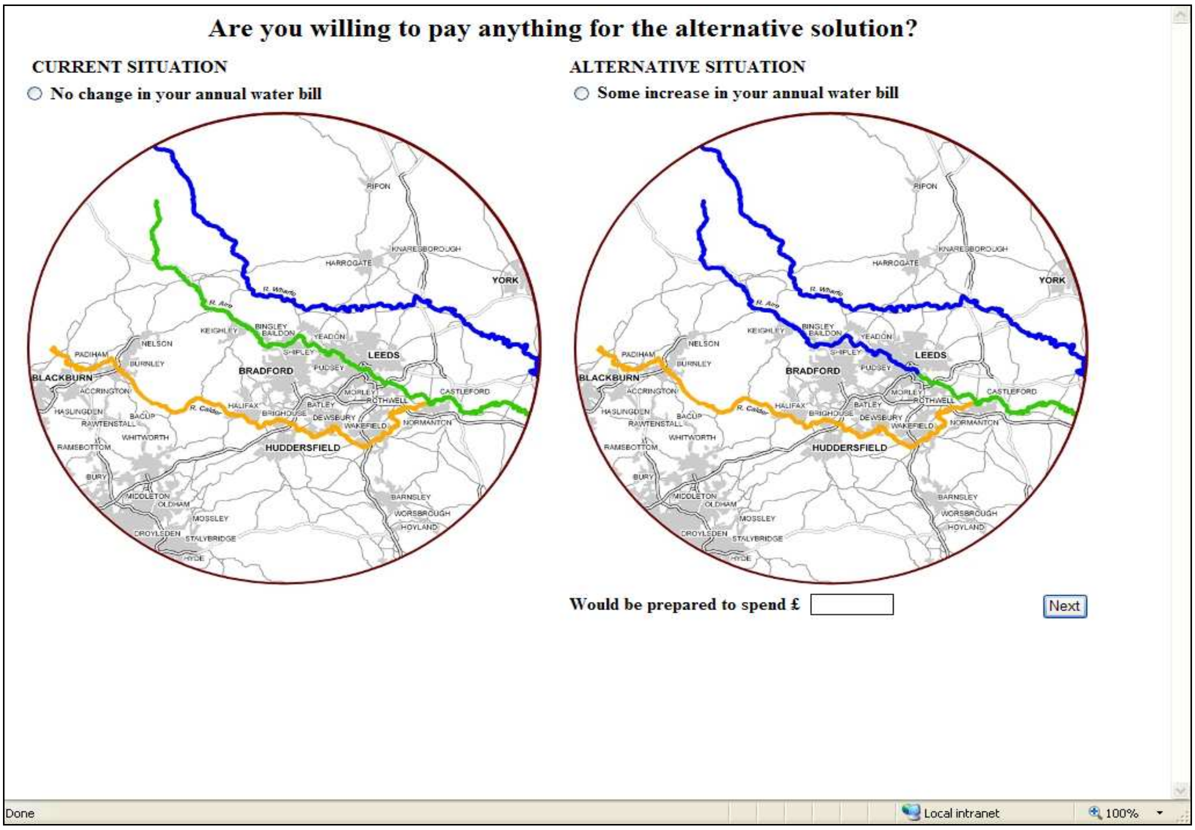
I now want to show you a second map [POINT TO “ALTERNATIVE” MAP]. This shows an alternative situation, where river water treatment works are undertaken to improve the stretch of river shown here [INDICATE CHANGED STRETCH]. Comparing the two maps you can see that in this stretch the river water quality has improved from YELLOW to BLUE.

We can see that’s a move from here [INDICATE INITIAL QUALITY] to here [INDICATE FINAL QUALITY]. All other parts of all the rivers stay as they currently are.

In a moment I will ask you a question about how much if anything your household might pay in increased water bills for this improvement. But before that please consider that any money you spend on improving river water quality obviously would not be available for spending on any other purchases. Please think about the location of the improvement, how close it is to your home, and whether you would benefit from it.

To help you work out how much, if anything, this scheme is worth to your household please consider this card. [GIVE RESPONDENT *PAYMENT CARD*]. For each amount please ask yourself whether or not your household would be prepared to pay this amount each year to get the improvement shown. Then tell me the amount which is the most your household would be prepared to pay on top of your normal yearly water bill in order to get this improvement.

Figure A.2: Screen shot – Valuation response question using payment card elicitation method



If Payment card value = 0

REASON A

E3. Looking at this list, please tell me the two most important reasons for your answer?

If Payment card value > 0

REASON B

E3. Looking at this list, please tell me the two most important reasons for your answer?

Now I would like you to consider a second alternative.

Again this concerns an improvement from [INDICATE] YELLOW to BLUE quality but now for this stretch of the river.

As before tell me the amount which is the most your household would be prepared to pay on top of your normal yearly water bill in order to get this improvement

Appendix: Payment card

Given that there is a clear literature showing that changes in the elicitation method used to pose WTP questions have significant impacts upon responses (Bateman et al., 1995; Bateman and Jones, 2003), this was standardised across all case studies using the common payment card[27] illustrated in Figure A3 which was prefixed by a standard budget constraint reminder. Although presented in local currency units, when converted into Euros the payment card included the same amounts for all countries. The payment card amounts were chosen after considering the differences in purchasing power between countries and the impact upon the statistical efficiency of WTP estimates of different payment card levels.

Figure A3: Common payment card (converted to Euro equivalents)

€0	€3	€6	€100	€135	€190	€350	€700	€1050
€3	€3	€7	€105	€140	€200	€400	€750	€1100
€5	€4	€7	€110	€145	€225	€450	€800	€1150
€1	€4	€8	€115	€150	€250	€500	€850	€1200
€1	€5	€8	€120	€160	€275	€550	€900	>
€2	€5	€9	€125	€170	€300	€600	€950	Other: € …..
€2	€6	€9	€130	€180	€325	€650	€1000	Don't know

A follow up question sought respondents' motivations for their WTP response which also allowed assessment of any protest responses, rejecting the valuation scenario (Bateman et al., 2002).

The UK case study additionally employed a dichotomous choice WTP elicitation approach with a separate sample of respondents.

[27] Note that the UK study also included a separate sub-sample for whom values were elicited using a single bound dichotomous choice elicitation method as recommended by Carson and Groves (2007). Results for this exercise are not presented within the present paper as elicitation effects preclude ready comparison with data from the other studies considered here.

Appendix: Procedural invariance testing

As pointed out in the main text, analyses of scope sensitivity are a necessary but often insufficient test for study validity. We emphasise the guidance obtained from experimental economics for such tests. Here the focus is upon whether findings pass tests of procedural invariance or exhibit anomalies. Our previous work provides a number of examples of such tests[28] (see, for example;

We can now formalise our procedural invariance test as follows defining

A^1 = Small quantity improvement, for which willingness-to- pay is denoted WTP (A^1)

and

A^2 = Large quantity improvement, for which willingness-to-pay is WTP (A^2)

To ensure that our test is not undermined by quality differences, A^2 is defined so that it contains all of A^1 plus an additional quantity of the good (i.e. A^1 is 'nested' within A^2). Therefore in both quantity and quality terms, $A^2 > A^1$. By varying the order of presentation randomly across respondents and denoting the 1st and 2nd question by subscripts, we therefore define the following four improvements over the status quo:

A_1^1, A_2^1, A_1^2, A_2^2

and their corresponding WTP measures:

[28] See, for example, Bateman, I.J., Burgess, D., Hutchinson, W.G. and Matthews, D.I., (2008) Contrasting NOAA guidelines with Learning Design Contingent Valuation (LDCV): Preference learning versus coherent arbitrariness, *Journal of Environmental Economics and Management*, 55: 127–141. http://dx.doi.org/10.1016/j.jeem.2007.08.003; Bateman, I.J., Cole, M., Cooper, P., Georgiou, S., Hadley, D. and Poe, G.L., (2004) On visible choice sets and scope sensitivity, *Journal of Environmental Economics and Management*, 47: 71-93. DOI: 10.1016/S0095-0696(03)00057-3; Bateman, I.J., Day, B.H., Dupont, D. and Georgiou, S., (forthcoming) Procedural invariance testing of the one-and-one-half-bound dichotomous choice elicitation method, *Review of Economics and Statistics*, in press; Bateman, I.J., Day, B.H., Jones, A. P. and Jude, S. (2009) Reducing gains/loss asymmetry: A virtual reality choice experiment (VRCE) valuing land use change, *Journal of Environmental Economics and Management*, 58: 106-118, doi:10.1016/j.jeem.2008.05.003; Bateman, I.J., Day, B.H., Loomes, G. and Sugden, R., (2007) Can ranking techniques elicit robust values? *Journal of Risk and Uncertainty*, 34:49–66, DOI 10.1007/s11166-006-9003-4; Bateman, I.J. and Langford, I.H. (1997) Budget constraint, temporal and ordering effects in contingent valuation studies, *Environment and Planning A*, 29(7): 1215-1228; Bateman, I.J., Langford, I.H., Jones, A.P. and Kerr, G.N. (2001) Bound and path effects in multiple-bound dichotomous choice contingent valuation, *Resource and Energy Economics*, 23(3): 191-213; Bateman, I.J., Langford, I.H., Turner, R.K., Willis, K.G. and Garrod, G.D. (1995) Elicitation and truncation effects in contingent valuation studies, *Ecological Economics*, 12(2):161-179. DOI: 10.1016/0921-8009(94)00044-V; Bateman, I.J. and Mawby, J. (2004) First impressions count: A study of the interaction of interviewer appearance and information effects in contingent valuation studies, *Ecological Economics*, 49(1): 47-55; Bateman, I.J. and Munro, A. (2005) An experiment on risky choice amongst households, *Economic Journal*, 115(502) March 2005: C176-C189. doi:10.1111/j.0013-0133.2005.00986.x; Bateman, I.J., Munro, A. and Poe, G.L. (2008) Asymmetric Dominance Effects in Choice Experiments and Contingent Valuation, *Land Economics*, 84: 115 - 127. http://le.uwpress.org/cgi/reprint/84/1/115; Bateman, I.J., Munro, A., Rhodes, B., Starmer, C. and Sugden, R. (1997a) A test of the theory of reference-dependent preferences, *Quarterly Journal of Economics*, 112(2): 479-505; Bateman, I.J., Munro, A., Rhodes, B., Starmer, C. and Sugden, R. (1997b) Does part-whole bias exist? An experimental investigation, *Economic Journal*, 107(441): 322-332; Covey, J., Loomes, G. and Bateman, I.J., (2007) Valuing risk reductions: Testing for range biases in payment card and random card sorting methods, *Journal of Environmental Planning and Management*, 50(4): 467-482. DOI: 10.1080/09640560701401986; Powe, N.A. and Bateman, I.J., (2003) Ordering effects in nested 'top-down' and 'bottom-up' contingent valuation designs, *Ecological Economics*, 45: 255-270.

$WTP(A_1^1)$, $WTP(A_2^1)$, $WTP(A_1^2)$, $WTP(A_2^2)$

We can now define a series of both scope sensitivity and procedural invariance tests. Because of the diminishing marginal utility associated with many environmental goods the utility of A^2 is not necessarily greater than A^1 utility[29]. Therefore combining a weak scope sensitivity expectation with our procedural invariance expectation that (within an exclusive list format) WTP for a given good should not vary by order of presentation:

$$WTP(A_1^1) = WTP(A_2^1) \leq WTP(A_1^2) = WTP(A_2^2) \quad (3)$$

while procedural invariance with strong scope sensitivity implies that:

$$WTP(A_1^1) = WTP(A_2^1) < WTP(A_1^2) = WTP(A_2^2) \quad (3a)$$

Furthermore we can also test for the consistency of scope sensitivity across question orders by defining:

$$A_2^2 - A_1^1 = \Delta BU$$

and

$$A_2^1 - A_1^2 = \Delta TD$$

where BU denotes the bottom-up ordering (smaller good (A_1^1) valued before the larger good (A_2^2)) and TD denotes the top-down ordering (larger good (A_1^2) valued before the smaller good (A_2^1)). These provide two estimates of the magnitude of scope sensitivity and preference consistency would lead us to expect that:

$$\Delta BU = \Delta TD \quad (4)$$

Results: Non-parametric scope sensitivity and procedural invariance testing

Table A2 provides an initial inspection of the scope sensitivity and procedural invariance results by apportioning each sample to a set of mutually exclusive response types. This analysis is revealing, showing that only 26% of the sample exhibit strong scope sensitivity while 45% accord equal values to each improvement. Of course the latter result is perfectly in accord with prior expectations, suggesting satiation at the smaller improvement level (which seems plausible given the nature of the good). Equally reasonable is that a further 25% of the sample accord no value to either improvement. Given the closer proximity of substitute goods mentioned above this again seems highly plausible. This leaves some 4% of responses that strictly clash with prior expectations in that the smaller improvement is accorded a higher value than the larger good. While this is an issue to be highlighted (and is often not tested for in non-market valuation studies) the rate of apparent irrationality or misunderstanding of the scenario is consistent with findings in experimental economic tests. While some studies have omitted data from such respondents we argue that this may give a misleading indication of the consistency and validity of findings and so retain all responses within subsequent analyses.

[29] Note that Hsee (1996) shows that a further anomaly can arise when a high quality good is given a higher preference rating than the same good plus an additional inferior (but still in its right utility enhancing) good. List (2002) shows that the same result is replicated within an incentive compatible, real payment framework.

Table A2: Classification of WTP response behaviour (sample percentages)

	Belgium	Denmark	Lithuania	Norway	UK	*Total*
WTP (A_1^1) < WTP (A_2^2)	4	23	5	9	14	*11*
WTP (A_1^2) > WTP (A_2^1)	10	25	6	15	20	*15*
WTP (A_1^1)> WTP (A_2^2) or WTP (A_2^1)< WTP (A_1^2)	6	2	2	7	1	*4*
WTP (A^1) = WTP (A^2)	72	33	27	44	39	*45*
WTP (A^1) = WTP (A^2) =0	8	17	60	25	25	*25*
	100	*100*	*100*	*100*	*100*	*100*

Resultant WTP levels disaggregated by both the size of improvement and the order in which valuations were sought are illustrated in Figure A4. A visual inspection suggests that, while each ordering treatment appears to yield scope sensitive results (with the larger improvement being accorded higher WTP) there appear to be considerable differences across the two treatments with greater scope sensitivity in the top-down treatment and a substantially higher WTP for the smaller improvement when presented as the first good encountered by respondents[30].

[30] This pattern accords with the previous findings of Bateman et al., (2004).

Figure A4: Scope sensitivity across ordering treatments

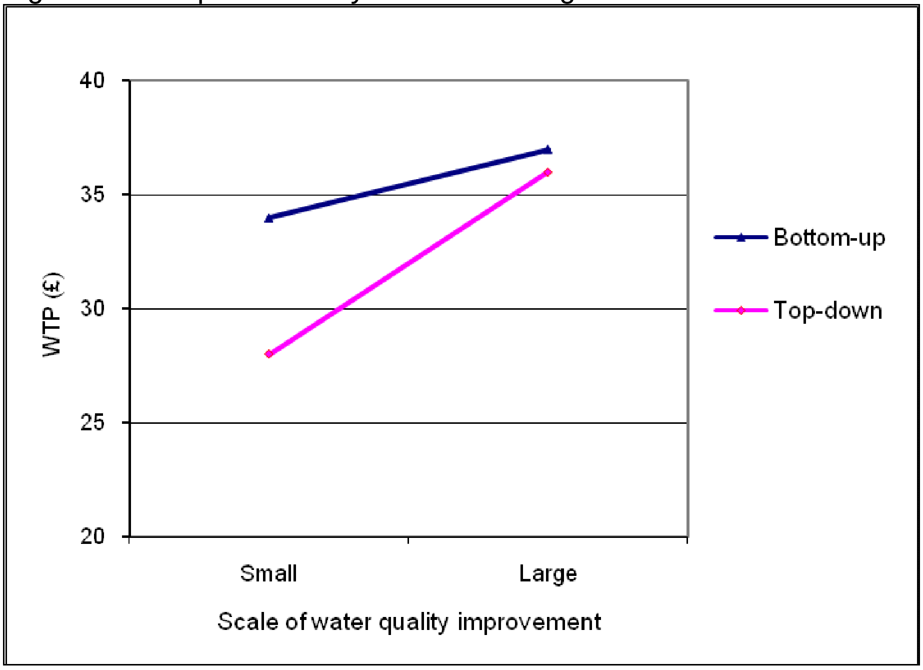


Figure A4 suggests that our WTP responses may reflect scope sensitivity but lack procedural invariance. To assess this, the expectations discussed in the main text were formulated into a series of testable hypotheses as set out in Table A3, which also reports results from nonparametric Kruskal-Wallis tests.

Table A3: Results from tests of scope effects and procedural invariance

Country	Scope effects				Procedural invariance (ordering effects)	
	All responses H_0^1: Small=Large	First WTP H_0^2: Small1=Large1	Second WTP H_0^3: Small2=Large2	Consistency H_0^4: ΔBU=ΔTD	Small good H_0^5: Small1=Small2	Large good H_0^6: Large1=Large2
Belgium	NS	NS	NS	S	S	NS
Lithuania	NS	NS	S	NS	NS	NS
Denmark	S	NS	S	NS	S	NS
Norway	S	NS	S	S	S	NS
UK	S	NS	S	S	S	NS

S = significant difference (α = 5%); NS = non-significant difference

Examining Table A3, H_0^1 (Small=Large) pools all WTP responses for the small improvement (irrespective of the order in which they were elicited) and compares these with pooled valuations of the large improvement. As argued previously, the only unambiguous expectation here is that WTP should not decline as the scale of the improvement rises. In the event all but two countries show significant increases in WTP as scope rises. As suggested, this is a relatively weak test but there are no anomalous reversals in scope sensitivity. This underlying weakness affects H_0^2 and H_0^3 which again report no anomalous scope reversals. However, it is interesting to note that when the first responses are tested (H_0^2: Small1=Large1) none of the country studies report significant scope sensitivity[31] whereas, by contrast, when the second responses are tested (H_0^3: Small2=Large2) all bar one yield significant scope effects. This indicates a worrying lack of scope consistency which is confirmed by H_0^4 (ΔBU=ΔTD) which reveals significant differences in the degree of scope across orderings in three out of five countries. These results suggest the presence of framing effects, which are further assessed within our procedural invariance tests H_0^5 and H_0^6. Hypothesis H_0^5 tests whether values for the small improvement are robust against whether responses were elicited from either the first or second valuation question. Results show that in only one case do we fail to reject this hypothesis; clearly procedural invariance fails for these values. However, when repeating this test for valuations of the large improvement we cannot reject the hypothesis (H_0^6) of procedural invariance in any country. This pattern prompts a number of speculations regarding response behaviour, one being that respondents may be overvaluing the small good when it is the first they encounter and they are unaware that alternative goods are also available. However, other interpretations are also plausible (some of which we develop further in Bateman et al., 2004) and we merely conclude that, as per the few other studies that have carried out such tests, we have found evidence of framing effects. Accordingly we account for the presence of ordering throughout our mean and value function transfer analyses.

[31] Note that this runs contrary to an incentive compatibility argument that first responses should give an unbiased and robust estimate of true WTP, devoid of strategic behaviour. Note however that this does not imply rejection of superficially similar argument of Carson and Groves (2007) as the latter only applies to responses to single referendum elicitation formats.

Appendix: Further specifications of the value transfer function.

In the main text of this paper we present three value function specifications for transfer purposes:

- The reduced theory-driven model: WTP = f(provision change, distance to site and distance to substitute)
- The full theory-driven model: WTP = f(provision change, distance to site and distance to substitute and income)
- Statistically-driven model: WTP = f(provision change, distance to site, distance to substitute, income, age and urban)

For completeness, Table A4 presents two further models:

- Best single variable theory-driven model: WTP = f(income). This is of interest because income is the strongest determinant of dissimilarity amongst our set of study countries. Further testing confirms that this provides the lowest function transfer error rate of any single variable model.
- The best-fit model: WTP = f(provision change, distance to site, distance to substitute, income, age, urban, user and visit frequency). Although this yields the highest level of explanation of study site data it was not selected for discussion in the main paper as the variables ‘user’ and ‘visit frequency’ are not available from secondary sources but would have to be elicited through a survey of the policy site, thus defeating the objective of benefit transfers which is to avoid such additional survey work.

Table A4: Results from further value function specifications estimated using random effects Tobit panel data methods.

	Best single variable theory-driven model	Best fit model
Variable	Coefficient (s.e.)	Coefficient (s.e.)
Constant	-4.63 (4.528)	-12.05 (10.289)
Large improvement	10.94 (0.989)	9.46 (1.105)
Small improvement 1st	6.050 (1.371)	4.87 (1.442)
Income (net household income in € per year)	0.0007 (0.0002)	0.0007 (0.0001)
Distance to the improvement site (in km)		-0.20 (0.078)
Substitute distance (distance to the nearest comparable site in km)		0.13 (0.038)
Age of respondent (in years)		-0.32 (0.102)
Urban (respondent lives in urban area=1; otherwise =0)		9.89 (3.346)
Users-non users (respondents visited one or more rivers or lakes in the last 12 months=1; otherwise=0)		15.65 (6.175)

Number of river and lakes recreation trips		0.037 (0.016)
Norway	11.51 (4.162)	18.69 (4.626)
UK	-15.70 (5.050)	-12.35 (5.47)
Belgium	22.18 (4.695)	31.94 (5.332)
Lithuania	-53.36 (5.447)	-45.79 (5.853)
Sigma μ	73.62	73.68
Sigma ε	23.31	23.31
Rho	0.91	0.91
No. of observations	5769	5268
Number of censored observations	1561	1430
Log-Likelihood	-23236	-21186
Wald chi^2(*K*=restriction for overall significance)	498 (7)	578 (13)
p-value	0.0000	0.000

Notes:

- The country dummy for Denmark is omitted making this the baseline from which country departures are to be interpreted
- The good/order dummy Small improvement 2nd is omitted making this the baseline from which the Large improvement 1st, Large improvement 2nd and Small improvement 1st departures are to be interpreted
- Rho = var_μ/(var_μ+var_ε) and represents the percent contribution to total variance of the panel-level variance component.

Appendix: Comparison of value transfer estimates with values predicted from policy site data.

In the main text we follow the approach of Van den Berg et al., (2001) and Brouwer and Bateman (2005) by comparing the value estimated by function transfer with the simple mean WTP of the policy site. However, some other analysts compare the transferred value with one estimated on data from the policy site alone (e.g. Barton 2002; Chattopadhyay 2003). Here we assess the transfer error by first defining the transferred value as $\hat{WTP}_{k|s}$ for policy site *k* estimated from study sites *s* and the directly predicted policy site value as $\hat{WTP}_k$. The transfer error is then calculated as $\left(\frac{\hat{WTP}_{UK|s} - \hat{WTP}_{UK}}{\hat{WTP}_{UK}}\right)\%$. Table A5 contrasts the simple mean WTP values with the directly predicted policy site values $\hat{WTP}_k$.

Table A5: Policy site sample mean and directly predicted WTP values.

	Lithuania	Belgium	Denmark	Norway	UK
	Mean WTP				
Small 1st	6	43	21	38	16
Small 2nd	6	50	29	45	22
Large	8	48	36	47	26
	Predicted WTP				
	Reduced theory-driven model (Provision change, distance to site and distance to substitute)				
Small 1st	9	52	30	51	20
Small 2nd	10	56	30	55	23
Large	11	56	39	56	26
	Full theory-driven model (Provision change, distance to site and distance to substitute and income)				
Small 1st	11	57	33	54	24
Small 2nd	13	62	34	58	27
Large	14	62	42	59	30
	Statistically-driven model (Provision change, distance to site, distance to substitute, income, age and urban)				
Small 1st	8	53	28	51	20
Small 2nd	10	58	29	54	23
Large	11	58	37	56	26

Table A6 details for the full dataset (including the dissimilar Lithuanian site) function transfer errors calculated by comparing values estimated from study sites with the value predicted from the same functional form applied to the policy site data. Tobit adjustments for censoring are incorporated within these estimates. Comparison with the value function transfer errors compared to simple mean values (shown in the main text) shows that a similar pattern of errors across different functional forms.

Table A6: Transfer errors for value functions compared with policy site predicted values – data including dissimilar (Lithuanian) site.

	Lithuania	Belgium	Denmark	Norway	UK	Average
Reduced theory-driven model (Provision change, distance to site and distance to substitute)						
Small 1st	57	93	66	93	84	79 (74)
Small 2nd	80	98	77	97	96	90 (84)
Big	32	88	69	88	77	71 (68)
Average (Weighted)	56 (50)	93 (92)	71 (70)	92 (91)	86 (84)	80 (75)
Full theory-driven model (Provision change, distance to site and distance to substitute and income)						
Small 1st	68	79	70	84	59	71 (73)
Small 2nd	89	82	81	88	54	80 (81)
Big	49	74	73	80	51	66 (68)
Average (Weighted)	69 (64)	78 (77)	75 (74)	84 (83)	55 (54)	72 (74)
Statistically-driven model (Provision change, distance to site, distance to substitute, income, age and urban)						
Small 1st	61	84	93	100	73	82 (85)
Small 2nd	83	87	102	105	82	92 (94)
Big	34	79	86	95	64	72 (76)
Average (Weighted)	60 (53)	83 (82)	94 (92)	100 (99)	73 (71)	82 (85)

Table A7 details for the reduced dataset of similar countries (excluding the dissimilar Lithuanian site) function transfer errors calculated by comparing values estimated from study sites with the value predicted from the same functional form applied to the policy site data. Tobit adjustments for censoring are incorporated within these estimates. Again comparison with the value function transfer errors compared to simple mean values (shown in the main text) shows that a similar pattern of errors across different functional forms.

Table A7: Transfer errors for value functions compared with policy site predicted values – data excluding dissimilar (Lithuanian) site.

	Belgium	Denmark	Norway	UK	Average
Reduced theory-driven model (Provision change, distance to site and distance to substitute)					
Small 1st	93	86	91	83	88 (89)
Small2nd	98	101	95	97	98 (98)
Big	87	87	86	77	84 (85)
Average (Weighted)	93 (91)	91 (90)	91 (90)	86 (84)	90 (90)
Full theory-driven model (Provision change, distance to site and distance to substitute and income)					
Small 1st	83	70	85	62	75 (77)
Small2nd	86	82	89	69	82 (83)
Big	77	73	80	57	72 (74)
Average (Weighted)	82(81)	75(74)	85(83)	63(61)	76 (78)
Statistically-driven model (Provision change, distance to site, distance to substitute, income, age and urban)					
Small 1st	78	67	93	54	72 (81)
Small2nd	81	75	98	49	77 (76)
Big	71	64	87	41	66 (70)
Average (Weighted)	77 (75)	69 (68)	93 (91)	48 (46)	72 (76)