Environmental and Resource Economics (2005) 30: 279–286

Validity Tests of Benefit Transfer – Are We Performing the Wrong Tests?

DADI KRISTOFERSSON and STÅLE NAVRUD*
*Department of Economics and Resource Management, Agricultural University of Norway, P.O. Box 5033, 1432 As, Norway; *Author for correspondence (e-mail: stale.navrud@ior.nlh.no)*

Accepted 27 July 2004

Abstract. The validity of environmental benefit transfer has been tested on numerous occasions assuming classical null hypothesis of equality. We argue against this assumption on the basis of theory, which clearly indicates that environmental benefits should be assumed to vary from context to context. We suggest the use of equivalence testing as a more appropriate and a clear compliment to the shortcomings of classical tests. Equivalence tests test the null hypothesis of difference between the original and transferred value estimates. Equivalence tests also combine the concepts of statistical significance and policy significance into one test, by defining an acceptable transfer error prior to the validity test. The results of a published study on validity of benefit transfer are reversed when subjected to an equivalence test.

Key words: benefit transfer, contingent valuation, equivalence tests, validity

JEL classifications: Q26, C12

1. Introduction

The validity of environmental benefit transfer has been the subject of a number of studies, see for example Loomis (1992), Loomis et al. (1995), Downing and Ozuna (1996) Kirchhoff et al. (1997), Brouwer and Spaninks (1999) and Bergland et al. (2002). In all cases, the validity has been tested by stating a null hypothesis of no difference between an original study result and a benefit transfer estimate. Rejection of such a null hypothesis is interpreted as evidence against the validity of benefit transfer, and non-rejection as evidence for validity. The interpretation of non-rejection depends, however, on the validity of the assumption of equality and the quality of the testing methodology. If theory does not support equal values as the most plausible state of nature, non-rejection is weakened as evidence for valid benefit transfer. The evidence is also weakened if the methodology is poor, resulting in large uncertainty of the estimates.

Received theory of the value of environmental goods predicts that their value depends on prices, income, nature of the good, and availability of

complements and substitutes (Johansson 1987): the individual's willingness to pay (WTP) for an improvement in environmental quality can be defined using the indirect utility function as:

$$V(\mathbf{p}, I, \mathbf{Q}_0) = V(\mathbf{p}, I - WTP, \mathbf{Q}_1)$$

where $\mathbf{p}$ is the vector of prices for goods and services, I is the individual's income and $\mathbf{Q}_0$ and $\mathbf{Q}_1$ are vectors describing environmental quality before and after the change respectively. This model hardly supports valid benefit transfer as probable state of nature. Valid benefit transfer is only conceptually plausible if the functional form of the indirect utility function in two populations is the same, and if the vector of prices and the vectors describing environmental quality are identical. The same change in environmental quality is obviously not sufficient; the level of environmental quality before and after the change, would also have to be the same. Desvousges et al. (1992) discuss the consequences this has on applicability of benefit transfer. The challenge, as they point out, is to identify environmental goods that are sufficiently similar to allow valid benefit transfer. Only in exceptional cases when the same population values the same environmental good at different points in time, we can assume that equality is the most plausible state of nature. We argue that the heterogeneity of environmental goods has consequences for the methods with which validity of benefit transfer should be tested. If equal environmental values are not to be expected then we should show caution when concluding that they are. In classical testing this would be type II assumption error, not rejecting the false hypothesis that environmental values are equal. This source of error is mostly neglected in econometric analysis. A more reasonable null hypothesis for a benefits transfer test would, in most cases, be that environmental values differ. This is the hypothesis for which non-rejection would produce the result expected from theory and previous empirical research (see Brouwer (2000) for a review of transfer error in earlier studies). We would like to identify a test methodology that reverses the burden of proof from classical tests and controls the probability of falsely assuming equal environmental values.

We suggest to use equivalence tests, which have been used for a number of years in pharmaceutical research to test whether drugs have equivalent properties (Hauck and Anderson 1984; Schuirmann 1987). Equivalence tests have, to our knowledge, not been used previously in economic research. The literature on the use of equivalence tests is extensive; see for example Hauck and Anderson (1984), Schuirmann (1987), Welling et al. (1991) and Berger and Hsu (1996). For examples from other fields of research, see Stegner et al. (1996) on psychological research.

The null hypothesis of an equivalence test is that values are different. Only through rejection of the null hypothesis can one conclude that the values are equivalent. Equivalence tests demand the definition of an interval within

which values are regarded equivalent, hence are equivalence tests and not equality tests. The interval must be predetermined by taking into account the practical use of the test results, which might prove limiting in some cases. The agreed upon standard in pharmaceutical research is 20% of the compared value.[1] Benefit transfer is an estimation method that will always be inaccurate, and acceptable error is therefore a central issue. Some sort of standard would have to be created to determine what constitutes acceptable transfer error.

When the natural null hypothesis is that the values are different, we believe that equivalence tests have properties that make them a clearly better alternative than classical tests. Even when the natural null hypothesis is that the values are equal, equivalence tests are clear compliments to classical test since they strengthen conclusions about transferability by directly testing the size of the transfer error against what is considered acceptable.

In this paper, we will discuss three useful properties of equivalence tests when testing for validity of benefits transfer: (1) assuming difference as the null hypothesis, (2) more reliable conclusions about transferability, and (3) explicit incorporation of what constitutes a policy-significant difference in values.

2. Equivalence Tests

In equivalence testing one reverses the roles of the null hypothesis (H_0) and the alternative hypothesis (H_A), and, by testing a set of these reversed hypotheses, demonstrates equivalence with a predetermined significance level. Suppose that it has been determined that a difference is negligible or policy insignificant if its absolute value is no greater than a small positive value Δ. The null and alternative hypotheses for an equivalence test would then be

$$H_0: \quad D \leq -\Delta \text{ or } D \geq \Delta$$
$$H_A: -\Delta < D < \Delta$$

where D is the difference between the two willingness to pay (WTP) estimates under investigation. If we can reject the null hypothesis then we conclude that H_A is accepted, i.e. the two are equivalent. If the null hypothesis is not rejected, we can only say that H_A has not been accepted. This result could be further investigated by a classical test of equality.

Hauck and Anderson (1984) and Schuirmann (1987) showed that if a $1–2\alpha$ confidence interval lies entirely between $-\Delta$ and Δ, the null hypothesis of non-equivalence can be rejected in favour of equivalence at the α level of significance. The equivalence test is at the α level because it involves two one tailed tests, which jointly describe the $1–2\alpha$ level confidence interval. A simple version of an equivalence test is the two one-sided test (TOST). More powerful tests than the TOST exist; see e.g., Berger and Hsu (1996). The simplicity and

widespread application of the TOST in e.g., pharmaceutical research, makes it a good tool for demonstrating the merits of equivalence testing.[2] It involves conducting two one-sided t-tests at the α level of significance.

$$t_2 = \frac{D - \Delta}{\sqrt{\sigma_D^2}} \geq t_{1-\alpha} \text{ and } t_2 = \frac{\Delta - D}{\sqrt{\sigma_D^2}} \geq t_{1-\alpha}$$

where $t_{1-\alpha}$ is the t-value associated with the chosen significance level and degrees of freedom, σ_D^2 is the variance of the difference. Schuirmann (1987) conducted a study of the rejection region of the TOST. His results along with the 95% confidence interval of a classical t-test are reported in Figure 1.

We can see that increased variance has different effects on the results of the two tests. This is to be expected, since the null hypothesis of the two tests are negatives of each other; equality for the classical test and difference for the equivalence test. As variance increases, the less you can say about the state of nature and therefore the less likely you are to reject your null hypothesis. The results are much more appealing for the equivalence test than for non-rejection of a classical null hypothesis, given that one wants to

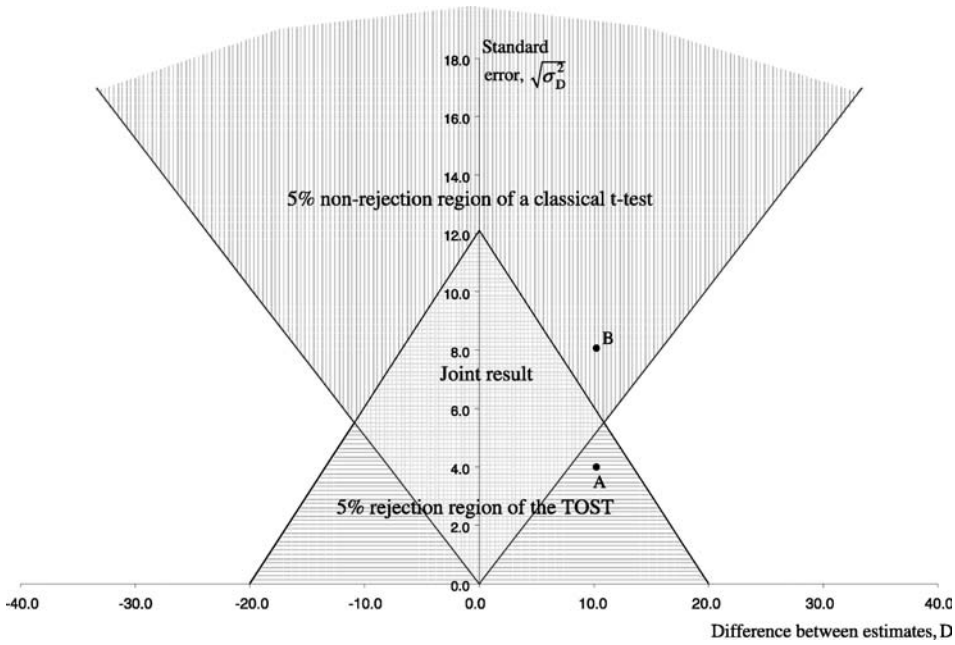

Figure 1. Comparisons of the rejection region for non-equivalence for the TOST (the upright triangle) and the non-rejection region for a hypothesis of equality tested by a classical t-test (the upside down triangle). D denotes the difference between the original estimate and benefit transfer estimate in absolute terms and $\sqrt{\sigma_D^2}$ denotes the associated standard error. For illustration of our point we assume a 5% level of significance and symmetrical acceptable transfer error of 20 % ($\Delta = 20$).

determine with certainty the validity of transfer. As variance increases, larger and larger transfer errors become valid according to the *t*-test. On the other hand, as variance increases it becomes increasingly difficult to reject the null hypothesis of difference for the equivalence test.

3. Examples

A simple numerical example will help to illustrate our point. Let us assume that two studies exist satisfying the general criteria of acceptable benefit transfer. Let us assume that an original estimate of WTP is 100 $/household/year and the benefit transfer estimate is 110 $/household/year. Let us assume that the standard error of the difference is 4 $/household/year. This benefit transfer would lead to a transfer error of 10%, which seems reasonably small. This transfer would however be rejected by a classical *t*-test at the 5% level of significance. This example is shown in Figure 1 as point A.

If the standard error is 8 $/household/year, the result is reversed and the transfer would not have been rejected by the *t*-test. This situation is shown in Figure 1 as point B. Increased variance in the estimates increases the likelihood that benefits transfer will be found to be valid.

If the TOST equivalence test procedure is used, and the 20% error margin generally used in pharmaceutical research is applied, then the results are reversed from that of the *t*-test. For example A, where the pooled standard error is 4 $/household/year, non-equivalence is rejected at a 5% level of significance, and we can assume equivalence of the two estimates. For example B, where the pooled standard error is 8 $/household/year, equivalence cannot be assumed at a 5% level of significance. To understand this result one must keep in mind the null hypothesis and what it says about our beliefs regarding the most plausible state of nature. The null hypothesis in equivalence testing is that the values are different. One wishes to show caution in concluding otherwise. One wishes therefore to be certain that a difference is not larger than a predetermined largest acceptable difference. Large variances make it difficult to determine the state of nature. Hence the result—increased variance decreases the likelihood that benefits transfer will be found to be valid.

We would like to support our results by applying equivalence tests to published results from benefit transfer validity studies. However, we were able to find only one study, Brouwer and Spaninks (1999), reporting the results on a detailed form that allow direct application of equivalence tests. They tested the validity of benefit transfer between meadowland sites in the Netherlands. The description of the compared meadowland sites did not clearly justify assuming that WTP values are equal. Still the authors applied a classical *t*-test to the data summarised in Table I, and found that their null hypothesis of equality was not rejected, although the difference in the two estimates was 27%. They concluded from this that the benefits transfer was "statistically valid."

Table I. Summary statistics from Brouwer and Spaninks (1999)

	Sample 1B	Sample 2
Untruncated mean	54.5	74.2
Standard error	11.3	5.8
Sample size	56	455

Their conclusion, regardless of the large transfer error, highlights the problems associated with use of non-rejection of a classical hypothesis to indicate equality. Given 27% transfer error, it is pointless to test for equivalence with a smaller expectable transfer error since it cannot be rejected for any significance level. However, to make a point, let us define a very large acceptable transfer error, e.g., 50%. Even at this level the equivalence test cannot reject the null hypothesis of difference. If it cannot be stated that one is 95% certain that the transfer error will be smaller than 50%, surely one must conclude that the values are different! The equivalence test does not support the original conclusion of Brouwer and Spaninks (1999) but instead concludes that difference is not rejected. The reason for the different conclusion is that the sample 1B is very small, and the estimated mean WTP has a large standard error. This results in a large standard error of the difference that puts us high on the *Y*-axis of Figure 1, above and to the right of point B. The results of Brouwer and Spaninks (1999) are in the area of Figure 1, where classical *t*-tests and equivalence tests tend to give contradictory results.

One major issue remains before equivalence tests can be adopted for testing the validity of benefit transfer, or any other econometric analysis. In order to assess the equivalence of two values, we must first define what would be considered equivalent. The agreed upon standard in the pharmaceutical industry is that the population mean tested must be within 20% of the mean of the reference group, as mentioned above. Such a standard must be set for each application. The standard must be based on what is considered policy relevant. An expert debate is necessary to determine acceptable transfer errors for different policy uses of estimates from benefit transfer exercises. The acceptable level does not have to be a single number but could vary depending on the use of the benefit transfer estimate. Navrud and Pruckner (1997) suggest that a higher level of accuracy is needed, and thus a lower level of acceptable transfer error, as we move from using benefit estimates in cost-benefit analysis for projects and policies, to environmental costing and green accounting and finally to natural resource damage assessments (NRDA) and other calculations of compensations for environmental damages from accidental or intended releases of polluting substances. Differences in legislation among countries could also lead to differences in acceptable transfer errors.

4. Conclusion

In this paper, we have argued that classical hypothesis testing has serious drawbacks when testing for the validity of environmental benefit transfer. Based on the theory of environmental valuation, we argue that difference is the more plausible null hypothesis in most cases. Researchers should take the consequence of this and use equivalence tests instead of the classical tests. The advantages of equivalence tests are that the null hypothesis is more in line with the prediction of theory. The probability of being able to conclude valid benefit transfer, given that it is the true state of nature, increases with more accurate estimates (i.e. smaller standard errors). Equivalence tests also combine the interpretation of results in terms of acceptable transfer error and the formal test within the test procedure.

Determining the appropriate null hypothesis is difficult and depends on a number of factors. We have argued for difference as a null hypothesis. It is however possible to construct scenarios where values should be expected to be equal, such as when the same population is asked to value the same good at two different points in time as in Downing and Ozuna (1996). Equivalence tests could therefore either be conducted in addition to classical testing, if the null hypothesis of equality is plausible, or as the primary tests if not.

We applied equivalence tests to published results where the authors concluded that the benefit transfer was valid. The original conclusion is not supported since difference cannot be rejected, even for extreme values of acceptable transfer error. In this case, significant additional information was attained about the results by conducting equivalence tests.

The issue of acceptable transfer error remains. It must be addressed before any widespread application of equivalence tests. We therefore call for a debate on the matter to determine what is considered acceptable, and whether one standard is enough or different standards must be defined for different types of policy use of benefit and damage estimates.

Acknowledgements

We would like to thank Richard C. Ready and two anonymous referees for very helpful comments.

Notes

1. The 20% standard has evolved over time much as the 5% level of significance is widely regarded as appropriate when valuing test results. This would clearly need to be addressed as equivalence tests start being used in benefit transfer.
2. However, results from CV studies based on maximum likelihood estimation of dichotomous choice or payment card interval data, are often reported as an empirical distribution

of WTP generated from Monte Carlo or bootstrapping techniques. This demands a more general equivalence test.

References

Berger, R. L. and J. C. Hsu (1996), 'Bioequivalence Trials, Intersection-Union Tests and Equivalence', *Statistical Science* **11**(4), 283–302.

Bergland, O., K. Magnussen and S. Navrud (2002), 'Benefit Transfer: Testing for Accuracy and Reliability', in R. J. G. M. Florax, P. Nijkamp and K. Willis, eds., *Comparative Environmental Economic Assessment*. Cheltenham, UK: Edward Elgar.

Brouwer, R. (2000), 'Environmental Value Transfer: State of the Art and Future Prospects', *Ecological Economics* **32**(1), 137–152.

Brouwer, R. and A. Spaninks (1999) 'The Validity of Environmental Benefit Transfer: Further Empirical Testing', *Environmental and resource economics* **14**, 95–117.

Desvousges, W. H., M. C. Naughton and G. R. Parsons (1992) 'Benefit Transfer-Conceptual Problems in Estimating Water-quality Benefits Using Existing Studies', *Water Resources Research* **28**(3), 675–683.

Downing, M. and T. Ozuna (1996), 'Testing the Reliability of the Benefit Transfer Approach', *Journal of Environmental Economics and Management* **30**, 316–322.

Hauck, W. W. and S. Anderson (1984), 'A New Statistical Procedure for Testing Equivalence in Two-Group Comparative Bioavailability Trials', *Journal of Pharmacokinetics and Biopharmaceutics* **12**(1), 83–91.

Hoenig, J. M. and D. M. Heisey (2001) 'The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis', *American Statistician* **55**(1), 19–24.

Johansson, P. O. (1987), *The Economic Theory and Measurement of Environmental Benefits*. Cambridge: Cambridge University Press.

Kirchhoff, S., B. G. Colby and J. T. LaFrance (1997), 'Evaluating the Performance of Benefit Transfer: An Empirical Inquiry', *Journal of Environmental Economics and Management* **33**(1), 75–93.

Loomis, J. B. (1992), 'The Evolution of a More Rigorous Approach to Benefit Transfer–Benefit Function Transfer', *Water Resources Research* **28**(3), 701–705.

Loomis, J., B. Roach, F. Ward and R. Ready (1995) 'Testing Transferability of Recreation Demand Models Across Regions – a Study of Corps of Engineer Reservoirs', *Water Resources Research* **31**(3), 721–730.

Navrud, S., and G. J. Pruckner (1997), 'Environmental Valuation – to Use or Not to Use? A Comparative Study of the United States and Europe', *Environmental and Resource Economics* **10**, 1–26.

Schuirmann, D. J. (1987), 'A Comparison of the Two One-Sided Procedure and the Power Approach for Assessing Equivalence of Average Bioavailability', *Journal of Pharmacokinetics and Biopharmaceutics* **15**(6), 657–680.

Stegner, B. L., A. G. Bostrom and T. K. Greenfield (1996), 'Equivalence Testing for Use in Psychosocial and Services Research: An Introduction with Examples', *Evaluation and Program Planning* **19**(3), 193–198.

Welling, P. G., F. L. S. Tse and S. V. Dighe (1991) *Pharmaceutical Bioequivalence*. New York: Marcel Dekker.