

Willingness to pay for insecticide treated nets and net re-treatment: An evaluation of Inter-rater and test-retest reliability.

Obinna Onwujekwe, Julia Fox-Rushby and Kara Hanson

Health Policy Unit, London School of Hygiene and Tropical Medicine, London WC1E 7HT

Abstract:

Objectives: To determine the inter-rater and test-retest reliability of willingness to pay (WTP) for insecticide-treated nets and net re-treatment using the bidding game (BG), binary with follow-up (BWFU) and a novel structured haggling technique (SH).

Methods: Inter-rater reliability was evaluated by having two sets of interviewers to administer questionnaires to 109 (BG), 110 (BWFU) and 103 (SH) randomly selected household heads. Test retest reliability was investigated by repeating interviews on 146 (BG), 161 (BWFU) and 139 (SH) household heads after one month of a first survey. The retest was accompanied by actual net sale. Data analysis used Pearson's correlation coefficient and 95% confidence intervals (C.I). The study was conducted in Achi community, Southeast Nigeria.

Findings: Inter-rater reliability coefficients were estimated for; whether the individual was willing to pay, WTP for own nets, WTP for others and WTP for re-treatment. Using WTP for own nets as the best reliability estimate, the coefficients were 0.77 (C.I: 0.72-0.86), 0.75 (C.I 0.64-0.81) and 0.74 (C.I: 0.63-0.82) in the BG, BWFU and SH respectively. There were six test-retest reliability coefficients. The coefficients for WTP for own nets were 0.51(C.I: 0.40-0.62), 0.41(C.I: 0.28-0.53) and 0.56(C.I: 0.41-0.65), for the BG, BWFU and SH groups respectively.

Conclusion: The findings show that the WTP method was reliable and that the elicitation methods had similar levels of reliability. The lower coefficients in the test-retest reliability could be due to the influence of factors affecting demand that had changed in the intervening period and also because the retest was accompanied by actual sale of nets. Standard formats for determining reliability within contingent valuation should be developed for easy comparison of results from different studies.

INTRODUCTION

Willingness to pay determination through the contingent valuation method is increasingly being used to generate information on the benefits and demand for health care programmes. The contingent valuation method is widely accepted as one theoretically method of estimating consumer surplus in order to get the full value of goods and services to consumers (Brookshire et al, 1980). The technique is also able to adapt to the special nature of health care (Golan and Shechter, 1993).

However, the reliability and validity of the contingent valuation method (CVM) has been called into question, and a long-standing criticism of the contingent valuation method is that stated willingness to pay (WTP) may be a poor indicator of actual WTP (Diamond and Hausman, 1994). A reliable measure is one that is relatively free from “measurement error”(Fink, 1993). Thus, “a contingent valuation study is reliable if unsystematic error is eliminated and valid if systematic bias is eliminated” (Mitchell and Carson, 1989).

Reliability concerns the extent to which an experiment, test, or any measuring procedure yields the same results on repeated trials (Carmines and Zeller, 1979). To enhance the reliability of a contingent valuation, the key scenario elements must be understandable, meaningful, and plausible (Mitchell and Carson, 1989). The problem of unreliable WTP estimates could stem from the type of elicitation method used, as Kealey and Turner (1993) found that individuals responded differently depending on the question format.

The reliability of the CVM has still to be proven (Klose, 1999). While most WTP studies have concentrated on determining the validity of the estimates there is a paucity of evidence on the reliability of the procedure. A systematic review found that only four studies in health care investigated test-retest reliability, with most studies using very small and unrepresentative samples (Klose, 1999) and the situation has not changed. The lack of evidence on reliability of WTP also extends to other applications of economics like environmental economics.

In both the health and other sectors, test-retest is the most common form of reliability determined. However, only O’Brien and Viramontes (1994), Flowers et al (1997) and Thompson et al (1984) could be said to be pure reliability studies since they presented reliability coefficients. The test-retest coefficients were 0.66 (O’Brien and Viramontes, 1994); 0.25 (Thompson et al, 1984) and 0.80 (Flowers et al, 1997).

In comparison to studies that have determined validity of WTP, those on reliability had been designed less elegantly and were more of ad-hoc studies. The sample sizes were also too low, 20 people for instance by O'Brien and Viramontes (1994). Some also used respondents that were unrepresentative of the general population like well educated people instead of using a mix as is found in population survey.

The analytic methods used to determine reliability were varied. As already seen, very few studies actually estimated reliability coefficients. Golan and Shechter (1993) used the general fit of their regression model as the criterion for stating that their result was reliable. The studies from other sectors namely Whitehead and Hoban (1999), and Desvosges et al (1993) used regression analyses and comparison of means to determine whether their estimates were reliable.

The time span between test and retest have ranged from one week to five years (Whitehead and Hoban, 1999). Loomis (1990) used a "retest" after nine-months and found that open-ended and close-ended WTP values were "reliable". Also Stevens et al (1994) found "reliable" test retest CV values after 3 years. However, it was difficult to decide whether the WTP estimates from these studies were strictly reliable or not since there were no objective measurements of reliability, as no coefficients or other test results were presented.

Compounding the picture of under-investigation of the reliability of the CVM is the use of research methods that do not measure reliability according to the conventional definition and characterisation in measurement theory. Hence, most of these studies that claimed to have investigated the reliability of the WTP method had not done so. Those that have used econometric models could be rightly said to have only looked at construct validity.

This paper therefore considers the test-retest and inter-rater reliability of the CVM using three elicitation methods; a novel WTP elicitation method¹ (that better mimics the price-taking mechanism in Southeast Nigerian); a bidding game; and the binary with follow-up format. The focus was the willingness to pay for insecticide-treated nets (ITNs) and the re-treatment of the nets for the control of malaria². ITNs programmes are new in Nigeria and people are expected to pay for them.

¹ The novel system is called the structured haggling technique

² The scenario and elicitation formats are available from the authors on request

METHODS

Context of research

The study was conducted in three malaria holo-endemic villages that form part of the twelve that comprise Achi autonomous community, which is located 5 kilometres from the Local government headquarters called Oji-River town and 45 kilometres from the State capital, Enugu. Achi is divided into two broad groups called Achi-uno and Achi-agu, because of links to common ancestors. There are six villages in each sub-division, with the three villages used in the study located in Achi-agu.

Achi is linked to Oji-River town by a single lane road covered with asphalt. However, the road is filled with potholes and presents formidable challenge to users during the rainy season. There are no tarred roads within the villages and dirt roads and bush paths provide means of access to the interiors. The Oji River runs very close to Achi, and actually forms one of the borders of two of the study villages. Ordinary mosquito nets are not sold in these villages, but they are sold in the urban towns like Enugu and Onitsha. It usually takes about one hour to reach Enugu and an extra 20 minutes to reach Onitsha by bus. This project was their first contact with ITNs.

The tests of reliability were part of a larger study determining the reliability and validity of different elicitation methods. The sample size for the broad study was 300 households per village. Thus, the samples used for both tests of reliability were sub-samples of those respondents. Respondents were household heads or their representatives (if the household head was not available).

The survey instrument

The questionnaire was divided into four parts. The first section was devoted to getting the personal data about the respondent and the household. In the next section, the WTP scenario and elicitation methods were administered. The scenario included showing the respondents a sample of the ITN. The last section was used to determine the households' asset holdings¹.

¹ The questionnaire is available from the authors on request

The survey for inter-rater reliability

Inter-rater reliability refers to the extent to which two or more individuals agree (Fink, 1993) about the responses/ratings of the same subject. The resulting intra-observer reliability coefficient measures the degree of variation between observers, with a score of 1.0 indicating complete agreement (Streiner and Norman, 1995).

Munro (1997) classified the strengths of the correlation coefficient (r) as: little, if any (0 – 0.25); low (0.26 - 0.49); moderate (0.50 - 0.69); high (0.70 - 0.89); and very high (0.90 - 1.00). Also O'Brien and Viramontes (1994) stated that their test-retest coefficient of 0.66 was acceptable, without stating the acceptability criterion.

The sample size was determined based on the reliability coefficients of previous studies using the bidding game². Thus, using the average of 2000 households in the three villages, at a 95% confidence level and a 80% power level, a minimum of 85 respondents was needed for the test of reliability. To allow for refusals and non-usable questionnaires, 150 respondents were selected in each village for the second survey in the test of inter-rater reliability, using systematic random sampling by including every 2nd household in the main sample of 300 respondents. The sample from each village was divided into three equal parts and an elicitation method applied to a third of each village respectively to control for differences in village characteristics that might affect the comparisons of the three elicitation techniques.

The pre-tested questionnaires were applied to the selected household heads or their representatives (if the household head was not available). The second interview by different interviewers using exactly the same questionnaire took place after a maximum of five days. This was because the respondents were either not available or needed time to recover from the first survey (it took an average of 30 minutes to complete) before submitting themselves to another.

² Out of all existing WTP elicitation techniques, the bidding game most resembles price-taking in the study area.

The survey for test-retest reliability

In test-retest reliability, the same test is given to the same people after a period of time (Carmines and Zeller, 1979). The usual test statistic is a correlation coefficient, and a correlation of 1.00 indicates the same results are obtained (Carmines and Zeller, 1979, Bowling, 1991). The interpretation of the coefficients is similar to that of inter-rater reliability.

In this study, the retest of the whole questionnaire (excepting questions about re-treating nets¹) was done within one month of completing the first survey with all respondents from that survey, with each respondent interviewed by the same interviewer. ITNs were offered for sale to all community members at a price of 350 Naira per net. The sellers of ITNs were the interviewers from the initial survey. They were instructed only to sell ITNs to people from households they had interviewed or from members of the community that were, not part of the survey sample.

If a respondent wanted to buy an ITN, a retest interview was requested prior to a purchase. If a member of a respondent's household wanted to buy an ITN, an interview with the original respondent was requested. If the person was at home² an interview was conducted (but the net was sold in advance of this interview) but if the person was not at home, the retest was recorded as missing. At the end of the net sale period, household interviews were held with respondents who had not made contact with the interviewers/ITN sellers and who agreed to be interviewed.

Data analysis

Friedman's test of related samples was used to determine the homogeneity between pairs of data. The test was also used to test the hypotheses that the WTP elicited by two interviewers (inter-rater) or at the two different periods (test-retest) respectively was the same.

Inter-rater and test-retest reliability were determined using Pearson's correlation coefficient. The variables examined for inter-rater reliability were; whether willing to pay a positive amount, level of WTP for ITNs for personal use, level of WTP for ITNs for other household members and level of WTP for re-treatment.

¹ This was because there was no time to follow-up the research with a re-treatment programme.

² Which would have been, on average, a 15 minute walk away

In the analysis of test-retest reliability, the WTP amounts determined during the second survey was not more than 350 Naira (the sale price of the nets). Thus, for comparing the WTP from the first survey and the re-test, the WTP amounts in the first survey that were more than 350 Naira were uniformly reduced to 350 Naira.

The reliability coefficients of the demographic and asset variables were also determined. This formed a check of the robustness of the WTP measurements because if the two sets of interviewers' differed considerably in measuring the demographic characteristics and assets of the households, then the robustness of the WTP reliability coefficients may be in doubt.

RESULTS

Inter-rater reliability

Demographic data of the respondents: 109, 110 and 102 respondents were interviewed in the bidding, BWFU and haggling groups (See Table 1). All groups were perfectly matched for sex and almost perfectly matched for the mean ages. The haggling group was perfectly matched for whether the respondents ever had any formal schooling, together with their marital status. The bidding and BWFU groups were also reasonably matched for marital status and whether the respondents had formal education. The number of years of formal schooling was most perfectly matched between the two interviewers in the bidding group followed by the BWFU group, and then the haggling group.

Table 2 shows the correlation coefficients of the same variables between raters. There were mixed findings, though all the coefficients were statistically significant at the 1% level. Gender was perfectly correlated between the pairs of interviewers in the three elicitation groups. The age, status in the household, whether the respondent had formal education and number of household residents was also highly correlated. However, while the number of years of schooling was highly correlated between raters in the bidding game and BWFU group, at 0.81 and 0.85 respectively, it was poorly correlated in the haggling group at 0.36. The inter-rater reliability correlation of the assets was mixed, particularly for ownership of a car, grinding machine and radio but there were no particularly discernible patterns across the elicitation methods.

Analysis of WTP by rater: Table 3 shows that the mean WTP amounts elicited by the first and second interviewers were much closer for WTP for own nets, than WTP for other household members' nets and for net re-treatment. The respondents' decisions of whether or not to pay anything for ITNs were statistically significantly different in the bidding group (chi-square 5.4; $p=0.02$). However, it was not statistically significantly different in the BWFU group (chi-square 0.50; $p=0.48$) and in the haggling group (chi-square 2.7; $p=0.10$).

There was no statistically significant difference in the mean WTP for own nets elicited by the pairs of interviewers in the three groups; Bidding (chi-square 1.29; $p=0.26$), BWFU (chi-square 0.14; $p=0.71$) and Haggling (chi-square 0.13; $p=0.72$). However, there were significant differences in the WTP for others in the Bidding game (Chi-square 7.11; $p=0.01$) and BWFU (chi-square 4.80; $p=0.03$), but not Haggling (chi-square=1.81; $p=0.18$). There was also a

statistically significantly difference in the bidding group (chi-square 6.43; $p=0.01$) for WTP for re-treatment although not so for the BWFU (chi-square 3.57; $p=0.06$) or haggling group (chi-square 1.8; $p=0.18$).

All the inter-rater correlation coefficients were statistically significant, based on a two-tailed test ($p<0.01$). The reliability coefficient on the decision whether or not to pay anything for ITNs was very high in all three groups. Table 4 shows the correlation coefficients, with the highest coefficient of 0.82 in the haggling group followed by 0.80 in the bidding game and 0.79 in the BWFU group. There was a similar high degree of correlation in the WTP for the respondents' own nets at 0.77, 0.75 and 0.74 in the bidding, BWFU and haggling groups respectively. However, the coefficients for other household members' nets were low to moderate; 0.58 (haggling), 0.54 (bidding) and 0.48 (BWFU). The coefficients for re-treating nets were moderate to, with the bidding game having the highest at 0.72, followed by the haggling group (0.59) and BWFU group (0.51). The confidence intervals around all the coefficients overlapped amongst the elicitation methods.

Test-Retest Reliability

Demographic data of the respondents: 146, 161 and 139 respondents were interviewed in the bidding, BWFU and haggling groups (see Table 5). The characteristics of the respondents reported by the two sets of interviewers like in the case of inter-rater reliability determination also showed greater measures of similarities. All groups were perfectly matched for sex and almost perfectly matched for the mean ages. The findings on holdings show that the results between the first survey and the retest were most uniform for ownership of a motorcycle, and with the other assets having varying degrees of similarity.

Table 6 shows that correlation coefficients for the same variables were all statistically significant at the 1% level. Gender was perfectly correlated across the three groups. All the other demographic variables showed test re-test reliabilities of more than 0.60, with the exception of status of the respondent in the household with a coefficient of only 0.45 in the bidding group. The findings on the assets were again mixed depending on the elicitation method, with the bidding group having the lowest coefficients and the BWFU have the largest number of higher coefficients.

Analysis of WTP over time: Table 8 shows that the mean WTP amounts elicited in the first and re-test surveys were much closer for WTP for own nets, than WTP for other household members' nets. The median WTP in both surveys were the same in the bidding and haggling groups, but different in the BWFU group. The respondents' decisions of whether or not to pay anything for ITNs were statistically significantly different in the bidding group (chi-square 4.80; $p=0.028$) and the BWFU (chi-square 4.48; $p=0.034$) but not in the haggling group (chi-square 3.24; $p=0.072$).

The respondents' WTP for own nets was not significant in the bidding (chi-square 3.57; $p=0.059$) or haggling group (chi-square 3.81; $p=0.051$), but was significantly different in the BWFU group (chi-square 18.68; $p=0.01$). Conversely, WTP for other peoples' nets was in the two periods was statistically significant in the bidding group (chi-square 17.89; $p=0.000$), but not in the BWFU (chi-square 0.28; $p=0.60$) and haggling group (chi-square 0.19; $p=0.66$).

Table 8 shows that all the test re-test reliability coefficients were statistically significant, based on a two-tailed test ($p<0.01$). The correlation coefficients were 0.56, 0.51 and 0.41 for WTP for own nets in the haggling, bidding and BWFU groups respectively. The test-retest coefficients for other household members' nets were lower than for own nets except for BWFU where was higher for others net. They were 0.52, 0.34 and 0.33 for the BWFU, bidding and the haggling groups respectively. The confidence intervals overlapped for WTP for own nets amongst the elicitation methods, but that for WTP for others was different in the BWFU group.

DISCUSSION

The analysis of demographic factors and asset holdings signalled that the interviewers had relatively few errors in either their questioning or recording of responses. The narrow margins of the confidence intervals for the WTP estimates were indicative of good precision and minimal presence of non-systematic errors. The findings of reliability are robust across the elicitation methods.

The contingent valuation method as a whole was a reliable technique for eliciting WTP. The strongest evidence was the inter-rater reliability using respondents' WTP for their own nets and the decision of whether or not to pay anything for ITNs, where all correlation coefficients were high. The reliability coefficients are comparable or higher than those generated by previous studies and the inter-rater reliability coefficients might have been higher if the period between the two interviews had not been as long as five days as happened in some instances.

Nonetheless, the case for the reliability of the CVM could be weakened by the lower test-retest coefficients. However, there could be genuine reasons for the lower coefficients, which may attest to the sensitivity and reliability of the CVM. As Whitehead and Hoban (1999) states, a CV instrument is still reliable if it measures significantly different WTP values across time periods, with differences in the appropriate direction, if any of the factors affecting WTP have changed over time. Thus a naïve interpretation of test-retest correlation can drastically underestimate the degree of reliability in measurement over time by interpreting true changes as measurement instability (Carmines and Zeller, 1979).

A major conceptual difficulty in establishing test-retest reliability is in determining how much time should lapse between the first and second administrations (Fink, 1993). If too much time elapses, external events might influence responses for the second administration; if too little time passes, the respondents may remember and simply repeat their answers from the first administration (Fink, 1993).

This study of inter-rater and test-retest could have suffered from both as respondents in both cases were re-interviewed after some days (inter-rater) and weeks (test-retest). In fact Whittington et al (1992) found that 16% - 20% of people who were given five to ten minutes to revise their WTP bids did so immediately and suggested that the short term reliability of CV studies through WTP could be questioned. However, it is most likely that the magnitude

of change or the influence of factors that will make the respondents change their minds would usually be less in inter-rater than in test-retest reliability because of the increased time difference between administration of the two surveys.

The fact that the re-test was accompanied by actual sale of nets could also have contributed in lowering the test-retest estimates in many ways. For instance, some people that really wanted to buy the nets but that previously stated a low WTP would readily pay the price of the net. However, they could have stuck to their earlier WTP if nets were not been sold. Conversely, some people that were not interested in buying the nets could have just stated very low amounts so as to be left alone.

In an attempt to minimise the influence of net sales, respondents were informed in the WTP scenario that nets would be sold after one month of the first survey. However, Mitchell and Carson (1989) suggest that hypothetical nature of contingent valuation scenarios introduces a random error without creating any systematic distortion and therefore would probably conclude that the actual sale of nets would be more likely to increase rather than reduce reliability.

The test re-test coefficient may also not be as high as inter-rater reliability coefficient because respondents may have changed their mind as a result of participating in the first survey (Bowling, 1991). This has been termed 'reactivity' by Carmines and Zeller (1979) who defined it to mean a situation where the process of measuring a phenomenon induces a change in the phenomenon itself. Whilst not usually related to this, the phenomenon would also be relevant to interpreting tests of inter-rater reliability.

The nature of value attributed to a good has an impact on the reliability of estimates of WTP. This was because the confidence intervals surrounding the reliability coefficients for respondents' WTP for their own nets were consistently higher than that for other peoples' nets. It is possible that respondents are either more knowledgeable about values for themselves than others, or that they thought less deeply about other people. Thus, they were less able to give similar answers when questioned a second time. Further support for this was found from the results of inter-rater reliability, where higher reliability coefficients were found across the three elicitation groups for WTP for goods for one's own household than for others.

It was not easy to classify the overall performance of the three elicitation methods because of the varied results. However, if the criterion was based on the respondents' WTP for own nets (because respondents are clearer about the value of a good for their own consumption), then the bidding and haggling methods were generally better than the BWFU method. The bidding method had the best inter-rater reliability and the haggling method generated the best test-retest data.

The coefficients of the demographic and asset holdings confirmed the robustness of the reliability coefficients, as they indicated that the interviewers were consistent in their measurements. The reliability coefficients for inter-rater reliability for the demographic and asset variables and WTP were all higher than the reliability coefficients for the demographic and asset variables and WTP for the test retest. This gives some indication that the reliability coefficients themselves are 'valid' indicators of reliability. If the test-retest coefficients had been low for the WTP questions but high for the demographic/asset questions, more doubt would have been placed on the reliability of the statements of WTP.

If the reliability coefficients for WTP estimates and the demographic/asset variables had been low, it might have indicated gross measurement error such that the data should be discarded. However, it might be possible to undertake further analysis to measure the direction of change in willingness to pay between the two surveys as a function of change in the dependent variables such that the higher the explanation of change, the greater the confidence in the WTP estimates.

CONCLUSION

This study provides evidence that the CVM could be used to generate reliable WTP estimates of the value attributed to a good for personal or household use in a mostly illiterate rural population in Nigeria. It is possible that such findings might be replicated in similar settings in other countries and in other parts of Nigeria. It is less clear that values attributed to consumption of a good by others is reliable and further work is needed. This is particularly important in the case of interventions to reduce infectious disease.

There needs to be a prioritisation within CVM studies of WTP to establish reliability, especially to examine test-retest in large enough samples. It could be argued that the most critical issue in the reliability of CVM is whether the same individual will elicit similar WTP values in repeated attempts. The next important issue is the extent to which two interviewers will elicit the same WTP estimate. This situation is important in surveys where some interviewers may not complete the interviews due to various reasons and the researchers do not want to lose the trend already established.

The interpretation of WTP reliability coefficients needs to be holistic and methods of analysis need to be directed towards resolving interpretations of test-retest results that account for changes in socio-economic and demographic variables over time. It is possible that this will improve with understandings of the factors that cause variation across individuals.

There needs to be greater consideration of the impact of the time interval between WTP surveys on the reliability of estimates, and an attempt to undertake estimates of inter-rater reliability with a shorter time interval. Finally, there should also be a standard definition of what reliability means in CVM by borrowing from measurement theory, since many studies have defined and measured reliability in ways that do not fit into any of the reliability categories. The methods of data analysis should also be standardised along the conventional definitions to increase comparability of results across settings.

Acknowledgements: This study received financial support from the UNDP/World Bank/WHO Special Programme for Research and Training in Tropical diseases.

References

- Bowling A (1997). *Measuring Health: A review of quality of life measurement scales*. (2nd edition). Buckingham; Open University Press:10-11. Carmines, E.G. and Zeller, R.A. (1979). *Reliability and validity assessment*. London; Sage Publications.
- Brookshire SD, Randall A and Stoll JR (1980). Valuing increments and decrements in natural resource service flows. *American Journal of Agricultural Economics*; 476-488.
- Carmines EG and Zeller RA (1979). *Reliability and validity assessment*. London; Sage Publications
- .
- Desvousges WH, Johnson RF, Dunford RW, Hudson SP, Wilson KN and Boyle KJ (1993). *Measuring natural resource damages with contingent valuation: Tests of validity and reliability*. London: Elsevier Publishers.
- Diamond P and Hausman J (1994). Contingent valuation: Is some number better than no number? *Journal of economic perspectives*;8(4): 45-64.
- Fink A (1993). *Evaluation Fundamentals: Guiding Health Programs, Research and Policy*. London: Sage Publications.
- Flowers, C.R., Garber, A.M., Bergen, M.R. and Lenert, L.A. (1997). Willingness to pay utility assessment: feasibility of use in normative patient decision support systems. *Journal of the American Medical Informatic Association Proceedings/AMIA Annual Fall Symposium*:223-227.
- Golan, E. H. and Shechter, M. (1993). Contingent valuation of supplemental health care in Israel. *Medical Decision Making*;13(4): 302-10.
- Kealey, M.J., and Turner, R.W. (1993). A test for the equality of close-ended and open-ended contingent valuations. *American journal of agricultural economics*; 75(2):321-331.
- Klose T.(1999). The contingent valuation method in health care. *Health Policy*; 47:97-123.

Loomis J.B. (1990) Comparative reliability of the dichotomous choice and open-ended contingent valuation techniques. *Journal of Environmental Economics and Management*; 18:78-85.

Mitchell, R.C. and Carson, R.T. (1989). *Using surveys to value public goods: the contingent valuation method*. Washington D.C.; Resources for the future.

Munro, B.H. (1997). *Statistical methods for health care research* (3rd edition). New York; Lippincott.

O'Brien, B. and Viramontes, J.L. (1994). Willingness to pay: A valid and reliable measure of health state preference? *Medical Decision Making*;14(3):289-97.

Stevens T.H., More, T.A. and Glass R.J. (1994). Interpretation and temporal stability of CV bids for wildlife existence: A panel study. *Land Economics*; 70:355-63.

Streiner, D.L. and Norman, G.R. (1995). *Health measurement scales. Practical Guide to their Development and Use*. Second edition. New York; Oxford University Press.

Thompson, M., Read, J. and Laing, M. (1984). Feasibility of willingness to pay measurement in chronic arthritis. *Medical Decision Making*; 4:195-215.

Whitehead, J.C. and Hoban, T.J. (1999). Temporal reliability in contingent valuation. *Land economics*;75(3):453-465.

Whittington, D., Smith, K., Okoroafor, A., Liu, J.L. and McPhail, A. (1992). Giving respondents time to think in contingent valuation studies: A developing country application. reliability of contingent values: A comparison of on-site users, off-site users, and non-users. *Journal of environmental economics and management*;22(3):205-225.

Table 1: Demographic data of the respondents for inter-rater reliability

Variables	Bidding		BWFU		Haggling	
	1 st Interviewer n (%)	2 nd Interviewer n (%)	1 st Interviewer n (%)	2 nd Interviewer n (%)	1 st Interviewer n (%)	2 nd Interviewer n (%)
Status						
1 = Head	76 (69.7)	83 (76.1)	85 (77.3)	89 (80.9)	73 (71.6)	81 (79.4)
0 = Rep	33 (30.3)	26 (23.9)	25 (22.7)	21 (19.1)	29 (28.4)	21 (20.6)
Total	109 (100)	109 (100)	110 (100)	110 (100)	102 (100)	102 (100)
Household residents: Mean (S.D)	4.19 (2.15)	4.28 (2.21)	4.46 (2.50)	3.91 (2.23)	4.80 (2.84)	4.34 (2.40)
Gender						
1 = Male	50 (45.9)	50 (45.9)	52 (47.3)	52 (47.3)	52 (51.0)	52 (51.0)
0 = Female	59 (54.1)	59 (54.1)	58 (52.7)	58 (52.7)	50 (49.0)	50 (49.0)
Total	109 (100)	109 (100)	110 (100)	110 (100)	102 (100)	102 (100)
Age (years): Mean (S.D.)	52.57 (16.43)	54.98 (15.99)	51.25 (14.47)	53.20 (16.40)	49.49 (14.52)	50.66 (14.29)
Attended school						
1 = Yes	51 (46.8)	56 (51.4)	62 (56.4)	66 (60.0)	63 (61.8)	63 (61.8)
0 = No	58 (53.2)	53 (48.6)	48 (44.6)	44 (40.0)	39 (38.2)	39 (38.2)
Total	109 (100)	109 (100)	110 (100)	110 (100)	102 (100)	102 (100)
Years of schooling: Mean (S.D)	2.80 (3.69)	2.91 (3.81)	3.87 (4.35)	4.00 (4.18)	4.47 (4.49)	5.38 (8.10)
Marital status						
1 = Ever married	96 (88.1)	98 (89.9)	94 (85.5)	92 (83.6)	88 (86.3)	88 (86.3)
0 = Not married	13 (11.9)	11 (10.1)	16 (14.5)	18 (16.4)	14 (13.7)	14 (13.7)
Total	109 (100)	109 (100)	110 (100)	110 (100)	102 (100)	103 (100)
Radio						
1 = Yes	80 (73.4)	67 (61.5)	87 (79.1)	87 (79.1)	66 (64.7)	67 (65.7)
0 = No	29 (26.6)	42 (38.5)	23 (21.9)	23 (20.9)	36 (35.3)	35 (34.3)
Total	109 (100)	109 (100)	110 (100)	110 (100)	102 (100)	102 (100)
Bicycle						
1 = Yes	69 (63.3)	74 (67.9)	69 (62.7)	69 (62.7)	61 (59.8)	59 (57.8)
0 = No	40 (36.7)	35 (32.1)	41 (37.3)	41 (37.3)	41 (40.2)	43 (42.2)
Total	109 (100)	109 (100)	110 (100)	110 (100)	102 (100)	102 (100)
Grinding machine						
1 = Yes	14 (12.8)	13 (11.9)	9 (8.2)	14 (12.7)	13 (12.7)	16 (15.7)
0 = No	95 (87.2)	96 (88.1)	101 (91.8)	96 (87.3)	89 (87.3)	86 (84.3)
Total	109 (100)	109 (100)	110 (100)	110 (100)	102 (100)	102 (100)
Motor cycle						
1 = Yes	18 (16.5)	19 (17.4)	28 (25.5)	28 (25.5)	17 (16.7)	23 (22.5)
0 = No	91 (83.5)	90 (82.6)	82 (74.5)	82 (74.5)	85 (83.3)	79 (77.5)
Total	109 (100)	109 (100)	110 (100)	110 (100)	102 (100)	102 (100)
Motor car						
1 = Yes	9 (8.3)	9 (8.3)	8 (7.3)	9 (8.2)	9 (8.8)	8 (7.8)
0 = No	100 (91.7)	100 (91.7)	102 (92.7)	100 (91.8)	93 (91.2)	94 (92.2)
Total	109 (100)	109 (100)	110 (100)	110 (100)	102 (100)	102 (100)

Table 2: Inter-Rater Reliability coefficients (demographic factors and assets)

	Bidding	BWFU	Haggling
Status	0.76	0.68	0.65
No. of residents	0.76	0.68	0.61
Gender	1.00	1.00	1.00
Age	0.75	0.87	0.91
Attended school	0.62	0.82	0.63
Years of schooling	0.81	0.85	0.36
Marital status	0.91	0.86	0.82
Own radio	0.55	0.45	0.72
Own bicycle	0.66	0.72	0.76
Own grinding machine	0.45	0.38	0.64
Own motorcycle	0.71	0.76	0.77
Own motorcar	0.68	0.30	0.68

All correlation coefficients were statistically significant at the 0.01 level (2-tailed).

Table 3: Willingness to pay elicitation for inter-rater reliability

Variables	Bidding		BWFU		Haggling	
	1 st Interviewer n (%)	2 nd Interviewer n (%)	1 st Interviewer n (%)	2 nd Interviewer n (%)	1 st Interviewer n (%)	2 nd Interviewer n (%)
Whether willing to pay						
1 = Yes	76 (69.7)	83 (76.1)	85 (77.3)	87 (79.1)	81 (78.6)	84 (82.4)
0 = No	33 (30.3)	26 (23.9)	25 (22.7)	23 (20.9)	22 (21.4)	19 (17.6)
Total	109 (100)	109 (100)	110 (100)	110 (100)	103 (100)	103 (100)
WTP for own nets						
Mean	161.10	170.00	193.68	192.91	192.28	188.74
(S.D)	(145.52)	(146.36)	(139.94)	(138.51)	(166.91)	(153.11)
Median	200.00	200.00	200.00	200.00	155.00	180.00
WTP for others						
Mean	93.59	57.43	92.59	71.68	89.40	120.06
(S.D)	(130.00)	(108.63)	(127.49)	(123.39)	(136.40)	(156.24)
WTP for re-treatment						
Mean	41.33	46.42	46.19	46.63	48.15	52.69
(S.D)	(30.69)	(33.73)	(21.05)	(22.99)	(28.76)	(30.57)
Median	30.00	50.00	50.00	50.00	50.00	50.00

Table 4: Inter-Rater Reliability coefficients for WTP

	Bidding	BWFU	Haggling
Whether WTP (95% C.I.)	0.80 (0.72 - 0.81)	0.79 (0.72 - 0.81)	0.82 (0.74 - 0.87)
WTP for own net (95% C.I.)	0.77 (0.72 - 0.86)	(0.75) (0.64 - 0.81)	0.74 (0.63 - 0.82)
WTP for others (95% C.I.)	0.54 (0.40 - 0.66)	0.48 (0.32 - 0.62)	0.58 (0.46 - 0.71)
WTP for re-treatment (95% C.I.)	0.72 (0.62 - 0.80)	0.51 (0.45 - 0.70)	0.59 (0.46 - 0.71)

All correlation coefficients were statistically significant at the 0.01 level (2-tailed).

Table 5: Demographic information for test-retest reliability

Variables	Bidding		BWFU		Haggling	
	1 st Interview n (%)	2 nd Interview n (%)	1 st Interview n (%)	2 nd Interview n (%)	1 st Interview n (%)	2 nd Interview n (%)
Status						
1 = Head	106 (72.6)	100 (68.5)	126 (78.3)	139 (86.3)	94 (67.6)	102 (73.4)
0 = Representative	40 (27.4)	46 (31.5)	35 (21.7)	22 (13.7)	45 (32.4)	37 (26.6)
Total	146 (100)	146 (100)	161 (100)	161 (100)	139 (100)	139 (100)
Gender						
1= Male	63 (43.2)	63 (43.2)	77 (47.8)	77 (47.8)	67 (48.2)	67 (48.2)
0 = Female	83 (56.8)	83 (56.8)	84 (52.2)	84 (52.2)	72 (51.8)	72 (51.8)
Total	146 (100)	146 (100)	161 (100)	161 (100)	139 (100)	139 (100)
Age (years): mean (S.D.)	51.88 (16.82)	50.45 (14.18)	50.05 (14.46)	50.40 (14.85)	48.61 (14.64)	50.53 (15.04)
Whether attended school						
1 = Yes	77 (52.7)	73 (50.0)	102 (63.4)	95 (59.0)	83 (59.7)	90 (64.7)
0 = No	69 (47.3)	73 (50.0)	59 (36.6)	66 (41.0)	56 (40.3)	49 (35.3)
Total	146 (100)	146 (100)	161 (100)	161 (100)	139 (100)	139 (100)
Marital status						
1 = Ever married	130 (89.0)	118 (80.8)	138 (85.7)	142 (88.2)	116 (83.5)	113 (81.3)
0 = Never married	16 (11.0)	28 (19.2)	23 (14.3)	19 (11.8)	23 (16.5)	26 (18.7)
Total	146 (100)	146 (100)	161 (100)	161 (100)	139 (100)	139 (100)
Radio						
1 = Yes	104 (71.2)	102 (69.9)	127 (78.9)	131 (81.4)	102 (73.4)	96 (69.1)
0 = No	42 (28.8)	44 (30.1)	34 (21.1)	30 (18.6)	37 (26.6)	43 (30.9)
Total	146 (100)	146 (100)	161 (100)	161 (100)	139 (100)	139 (100)
Bicycle						
1 = Yes	95 (65.1)	96 (65.8)	97 (60.2)	96 (59.6)	82 (59.0)	76 (54.7)
0 = No	51 (34.9)	50 (34.2)	64 (39.8)	65 (40.4)	57 (41.0)	63 (45.3)
Total	146 (100)	146 (100)	161 (100)	161 (100)	139 (100)	139 (100)
Grinding machine						
1 = Yes	16 (11.0)	12 (8.2)	17 (10.6)	23 (14.3)	29 (20.9)	23 (16.5)
0 = No	130 (89.0)	134 (91.8)	144 (89.4)	138 (85.7)	110 (79.1)	116 (83.5)
Total	146 (100)	146 (100)	161 (100)	161 (100)	139 (100)	139 (100)
Motor cycle						
1 = Yes	27 (18.5)	27 (18.5)	29 (18.0)	29 (18.0)	26 (18.7)	30 (21.6)
0 = No	119 (81.5)	119 (81.5)	132 (82.0)	132 (82.0)	113 (81.3)	109 (78.4)
Total	146 (100)	146 (100)	161 (100)	161 (100)	139 (100)	139 (100)
Motor car						
1 = Yes	12 (8.2)	18 (12.3)	12 (7.5)	15 (9.3)	13 (9.4)	9 (6.5)
0 = No	134 (91.8)	128 (87.7)	149 (92.5)	146 (90.7)	126 (90.6)	130 (93.5)
Total	146 (100)	146 (100)	161 (100)	161 (100)	139 (100)	139 (100)

Table 6: Test re-test Reliability coefficients (demographic factors and assets)

	Bidding	BWFU	Haggling
Status	0.45	0.76	0.80
Gender	1.00	1.00	1.00
Age	0.69	0.76	0.91
Attended school	0.62	0.81	0.62
Marital status	0.61	0.72	0.83
Own radio	0.37	0.65	0.44
Own bicycle	0.50	0.63	0.65
Own grinding machine	0.37	0.61	0.44
Own motorcycle	0.64	0.71	0.69
Own motorcar	0.49	0.56	0.52

All correlation coefficients were statistically significant at the 0.01 level (2-tailed).

Table 7: Willingness to pay elicitation for test-retest reliability

Variables	Bidding		BWFU		Haggling	
	1 st Interview	2 nd Interview	1 st Interview	2 nd Interview	1 st Interview	2 nd Interview
WTP for own net						
Mean	158.29	167.53	190.03	220.92	190.07	194.64
(S.D)	(109.09)	(133.86)	(107.47)	(135.36)	(126.97)	(145.72)
Median	200.00	200.00	200.00	250.00	200.00	200.00
WTP for others						
Mean	78.70	47.53	101.89	101.75	108.35	73.13
(S.D)	(99.65)	(91.22)	(121.94)	(129.39)	(168.74)	(126.94)

Table 8: WTP Test re-test Reliability coefficients

	Bidding	BWFU	Haggling
WTP for own net (95 C.I.)	0.51 (0.40 – 0.62)	0.41 (0.28 – 0.53)	0.56 (0.41 – 0.65)
WTP for others (95% C.I.)	0.34 (0.20 – 0.47)	0.52 (0.48 – 0.68)	0.33 (0.20 – 0.47)

All correlation coefficients were statistically significant at the 0.01 level (2-tailed).