Manuscript Details

Manuscript number	*ECOLEC_2016_789*
Title	*Stability of willingness-to-pay for coastal management: a choice experiment across three time periods*
Article type	*Research paper*

Abstract

A key assumption of stated preference methods is that individuals have well-formed preferences that are robust over time. Both the discovered and constructed preference perspectives imply this is not necessarily the case. There can be a large situational component to expressed preferences that add to the uncertainty of sampling error. Most non-market valuation studies only collect data from one point in time so the degree of temporal variability cannot be tested. Test-retest studies that provide data from two points in time generally find significant differences in preference structure and willingness-to-pay (WTP). In this study we test stability of WTP for beach erosion management using a fully ranked discrete choice experiment survey with not one but two retests over a six month period. We find that stability does not improve with the additional repetition as the preference discovery hypothesis implies it might. WTP confidence intervals overlap but the models are significantly different at each point in time, even after allowing for variation in choice error. Either the survey did not facilitate sufficient preference discovery or preferences were reconstructed. However, respondents with high scores of self-reported certainty in their choices in the first survey had significantly more stable WTP estimates.

Keywords	*Preference stability, choice experiment, coastal erosion, New Zealand*
Manuscript category	*Analysis*

Corresponding Author	*Yvonne Matthews*
Corresponding Author's Institution	*University of Waikato*
Order of Authors	*Yvonne Matthews, Riccardo Scarpa, Dan Marsh*

Submission Files Included in this PDF

File Name [File Type]

Sub 1 - Title Page.doc [Title page]

ECOLEC_2016_789 - Response to reviewers.docx [Response to reviewers]

ECOLEC_2016_789 - Sub 2.4 no markup.docx [Manuscript]

ECOLEC_2016_789 - Sub 2 - Highlights.docx [Highlights]

Stability of willingness-to-pay for coastal management: a choice experiment across three time periods

Author Details

Yvonne Phillips

1. Waikato Management School, University of Waikato, New Zealand

Riccardo Scarpa

1. Gibson Institute for Land, Food and Environment, and Institute for Global Food Security, Queens University Belfast, Northern Ireland
2. Waikato Management School, University of Waikato, New Zealand

Dan Marsh

1. Waikato Management School, University of Waikato, New Zealand

Address for correspondence

Department of Economics, University of Waikato, Private Bag 3105, Hamilton, New Zealand;

Phone: 64 21 976 169

Email: yvonneresearch@gmail.com

Acknowledgements

We are very grateful for research funding contributed by Waikato Regional Council and Waikato Management School

Response to reviews

Author summary

We are very grateful to both reviewers for their time and detailed feedback. We particularly appreciate the advice from review 1 to include a more in-depth discussion of discovered and constructed preferences and from reviewer 2 to use WTP-space models. We believe this revised paper is a significant improvement over the previous version. We respond to comments in detail below.

Comments from the editors and reviewers:
-Reviewer 1

This is a very good and interesting paper that addresses a relevant research question, i.e. whether or not willingness-to-pay estimates in choice experiments in the context of environmental valuation are stable. The methodology used, a longitudinal choice experiment over three periods of time, is relatively novel and suitable. Overall, the paper has a clear, easy-to-follow structure and is well-written.

Thank-you for these positive comments

I see one major weakness of the paper. The authors need to discuss the theoretical background of preference "construction" and similar concepts more thoroughly. I recommend to add an additional chapter after the introduction. It's important that the authors discuss different concepts, such as "preference construction/formation" and "preference discovery" in a structured and comprehensive way. How do both concepts explain choices in new, uncertain, complex situations? Both concepts are well-known, but quite different in terms of explaining the decision-making process and in terms of their implications for the validity of stated preferences. The additional chapter should also lead to and clearly state the research hypotheses which are later tested using different models and tests.

After reviewing more literature about discovered and constructed preferences we concur that these concepts are highly relevant, thank-you. We have added this section although it required some rearrangement to the rest of the paper to keep the total word count manageable.

Some further comments:
Page 2, Line 24-29: Do you disregard studies with repeated preference statements which occur within the same survey wave (e.g. repeated preference elicitation on the same day within a valuation workshop)?

Have added a comment to say these studies are still useful indicators of short-term reliability, although our focus is on the longer term.

Page 2, Line 33-34: Please clarify "reliability coefficients". What does a reliability coefficient of 0.3 mean?

Have elaborated to explain this means the correlation between responses at different points in time.

Page 3, Line 16-26: You may want to take into account Lienhoop, N., & Völker, M. (2016). Preference Refinement in Deliberative Choice Experiments for Ecosystem Service Valuation. Land Economics, 92(3), 555-577. Their study also includes a longitudinal choice experiment for ecosystem service valuation.

Thanks, we have read and included this now. As well as a new paper by Czajkowski et al (2016)

Page 3: At the end of the introduction, state clear research questions!

Done. New paragraph says:
"Our first research question is how stable is WTP in our specific context, and is this consistent with other test-retest studies? But the more interesting and unique question is does stability improve between the first and second re-test? If so, it would be consistent with the concept of learning and preference discovery. If not, the results would be more consistent with the transience of preferences constructed on the spot. We also investigate to what degree choice consistency can be explained by individual-specific factors. If preference stability could be predicted this could improve confidence in one-shot experiments where retest is not an option."

Page 4, Equation 1: It would be more consistent to include the time dimension here (just as in equation 2).

Done

Page 6, Line 5: Explain here or before what Coromandel is!

The end of the introduction now mentions that the study location is a peninsula named Coromandel

Page 6, Line 12-13: Did you test that somehow? E.g. by asking respondents about their choice strategy (e.g. "I just chose the same alternatives as last time.")? That would be a robust test for a potential memory effect.

Unfortunately we did not. I have added a statement to this effect.

Page 6, Line 26-27: "Differences in scale might be caused by learning or fatigue effects rather than result from changed preferences." - The authors need to be more precise here! Do preferences change due to learning or is learning something different from preference change? How is preference change caused then? See my comment above on "preference construction/formation" and "preference discovery". This is a complex issue which needs to be explained much better in this manuscript.

Scale is a complex issue and we don't have a lot of space to discuss the nuances but we hope the new section about discovered and constructed preferences addresses this to your satisfaction.

Page 8: How was the survey sample selected?

Added this to the Survey Instrument section:
"Respondents were selected from a pre-recruited panel of New Zealand residents provided by a market research company and a smaller, self-selected sample from online advertisements on Facebook and Google. To qualify for the survey, respondent had to have visited the peninsula in the previous twelve months."

Page 11, Line 3-6: Which test was used for that?

The new paragraph says:
"The lack of the significance of this variable combined with the fact that there is no overlap between significant variables for retest participation and those explaining choice congruency (**Error! Reference source not found.** in the appendix) implies consistency results are unlikely to be affected by selection bias"

Page 11, Line 11-12: Why would you expect that?

Added the explanation that 17% is equivalent to 1/6 (the number of alternatives)

Page 12, Figure 2: Is this the absolute difference in rank compared to wave 1? Please clarify!

Yes it is in comparison to wave 1. I've altered the axis label to say this

Page 12, Figure 2: Please add the number of respondents in the different waves!

Done

Did you (also in your other models and tests later on) analyze data from all respondents or only those respondents who answered in all three waves? I would favor using only those who answered in all three waves.

I admit the pairwise comparisons were motivated by reluctance to discard the data from people who only did two out of three waves. But you're right it is best to focus on people who did all three. The results section has been re-written to reflect this

Page 13-14, Table 4: How can the number of individuals in wave 3 (429) specified here be higher than the number of individuals in wage 3 (426) specified in Table 2?

Typo, sorry. 426 is correct

Page 19, Line 11-12: Elaborate more on that! Why do your results support the preference construction hypothesis? What would you conclude about the preference discovery hypothesis then? But is your evidence really sufficient? What could be alternative explanations for your results? Fatigue, strategic behavior, lack of information, etc.?

We have added an alternative explanation that the lack of feedback on choice consequences may have prevented any real preference discovery from happening.

Page 19, Line 24: Please clarify, what kind of things matter and what kind of things don't matter!

This sentence is now gone – replaced by the discovered/constructed preference discussion

Page 19, Line 27 ff.: Elaborate more on ways to maximize the likelihood of eliciting well-formed preferences. E.g. through use of deliberative choice experiments.

We now discuss the importance of feedback to preference discovery and in the conclusion mention that deliberative choice experiments might help. Alternatively, virtual reality or a personalized, hypothetical rates invoice might also help.

-Reviewer 2

-

Dear Authors,

you present an interesting paper implementing a test-retest analysis to evaluate WTP stability. I have a few major and minor comments.

Major comments

a) You need to do a better job at highlighting the original contribution of your paper. What does an additional round of re-testing really bring to the table, compared to previous studies?

We now discuss (page 4) that the additional re-test allows us to explore the discovered preference hypothesis that stability improves with repetition.

b) Section 2.3.1.1 illustrates your test of choice consistency. As you recognize, the test is very strict. There are six choices in each of your cards. If the preferences of an individual are 1,2,3,4,5,6 in the first survey and 1,2,3,4,6,5 in the second survey, your test will conclude that his/her preferences are not stable. In reality, I find the two choices very consistent. You should illustrate this issue in a much clearer way and identify an approach to deal with it.

We acknowledge that this could have been explained better in the method section. The restrictiveness of choice congruency is the reason why we also compare utility and WTP functions. Have added the following paragraph to section 2.3.1:

"Choice congruency is a rather restrictive test considering we ask respondents to fully rank six alternatives. It would be difficult for someone to rank them in exactly the same order each time. A change of one position is less inconsistent than a complete reversal of ranks so we also report the absolute difference in ranks for both waves. However, we do not report a linear regression for difference in rank because rank is an ordinal not a lineal variable and it is not strictly correct to treat it as such."

And this to the beginning of section 2.3.2:

"Testing for equality of the random utility function allows for random error in responses"

c) You report that WTP measures in your analysis have implausible high ranges. This is a well-known issue of models estimated in preference space (e.g. Scarpa et al., 2008). Since this is the focus of your comparison across surveys, you should really consider using a WTP space model. In my opinion, comparting those WTPs across surveys would be a much more interesting test of consistency across time then the ones you report.

Thank-you for the very useful suggestion. We tried the WTP-space model and found the overall fit was slightly worse but it did reduce the variance of WTP so decided to use it. It does make it easier to compare WTP across waves, although of course the WTP simulation is still required to approximate the sampling distribution.

Minor comments

a) In section 2.1 you talk about your RUM saying that you "assume the probability of a consumer choosing a beach destination is a function of …". You do not really do a travel cost model, but a choice experiment and your method section should reflect that. The current illustration is confusing.

Thanks for noticing this. We're also working on a destination choice study hence the confused writing. I have changed the text to:
"assume the probability of a consumer choosing their preferred future state of a beach is a function of…"

Also in the same section it appears that you are assuming all parameters to be random, while then in the application most of them are fixed. I would suggest re-writing this section to avoid this confusion.

It is true that the alternative specific constants are non-random. I have altered the equations to reflect this.

b) In section 2.2. you should explain better what you are testing with the binary logit. It becomes clear later on in the manuscript but at first it was a bit obscure.

I have added the following to section 2: "to test whether there is a strong relationship between demographic variables and re-test participation. For respondents who complete the first retest we also test whether choice consistency is a significant explanatory variable for participation in the second retest. "

c) Page 6 line 23. Please add "...explicit scale parameter -in the pooled model- ".
Done. This section has also been altered to due to the change to WTP-space models

References

Scarpa R., Thiene M., Train K. (2008) Utility in willingness to pay space: a tool to address confounding random scale effects in destination choice to the Alps, *American Journal of Agricultural Economics*, vol. 9: 994–1010.

Stability of willingness-to-pay for coastal management: a choice experiment across three time periods

Abstract

A key assumption of stated preference methods is that individuals have well-formed preferences that are robust over time. Both the discovered and constructed preference perspectives imply this is not necessarily the case. There can be a large situational component to expressed preferences that add to the uncertainty of sampling error. Most non-market valuation studies only collect data from one point in time so the degree of temporal variability cannot be tested. Test-retest studies that provide data from two points in time generally find significant differences in preference structure and willingness-to-pay (WTP). In this study we test stability of WTP for beach erosion management using a fully ranked discrete choice experiment survey with not one but two retests over a six month period. We find that stability does not improve with the additional repetition as the preference discovery hypothesis implies it might. WTP confidence intervals overlap but the models are significantly different at each point in time, even after allowing for variation in choice error. Either the survey did not facilitate sufficient preference discovery or preferences were reconstructed. However, respondents with high scores of self-reported certainty in their choices in the first survey had significantly more stable WTP estimates.

Keywords

Preference stability, choice experiment, coastal erosion management, New Zealand

1. Introduction

When using stated preference methods to learn about preferences for the environment we ask people to explore and state their willingness-to-pay (WTP) for hypothetical alternatives. An important issue in stated preference research is whether these hypothetical decisions are reliable. Results may be used today from studies conducted years ago in both policy design and benefit transfer. In these cases a fundamental maintained assumption is that these values are robust over time (Brouwer, 2006). It is important for decision makers and practitioners to know to what degree this is the case. Rational choice theory allows WTP to vary for reasons such as changes in the choice context or changes in individual circumstances. Individuals who gain new consumptive experience such as experiencing a change in environmental quality may alter their preferences (McConnell, Strand, & Valdés, 1998). But are preferences stable in the aggregate?

1.1 Evidence on stability

Discrete choice experiments (DCEs) allow explicit testing of the stability of the utility function and choice consistency. There does not appear to be any difference in reliability compared with other stated preference elicitation methods such as contingent valuation (Liebe, Meyerhoff, & Hartje, 2012). Some DCE studies use repeated choice questions within the same survey which provide clues about choice reliability in the very short term. Choices have been shown to vary over the duration of a single survey due to learning (about the choice task) or fatigue (Hess, Hensher, & Daly, 2012) , but in other cases due to strategies (Day et al 2012). Attrition is a major problem in longitudinal studies so most stated preference studies merely provide information from one point in time. Some use different samples (e.g. Bliem, Getzner, & Rodiga-Laßnig, 2012) but it is then impossible to control for unobservable sample differences. However, there are examples in the literature where a re-test was conducted either weeks or months after the original survey.

Several DCE studies report 60-80 percent congruent choices for retests within weeks or months of the first test in the area of health economics (Bryan, Gold, Sheldon, & Buxton, 2000; Ryan, Netten, Skåtun, & Smith, 2006; Skjoldborg, Lauridsen, & Junker, 2009) and food preferences (Carlsson, Mørkbak, & Olsen, 2012; Rigby & Burton, 2011). Unlike healthcare or food, environmental quality is typically a public good with components of non-market and non-use value and may have greater WTP variability (Carlsson, 2010). Bliem, Getzner and Rodiga-Laßnig (2012) report that WTP for water quality varied by up to 39 percent using two independent samples a year apart. Liebe, Meyehoff and Hartje (2012) find preferences for wind farms are significantly different after eleven months, but assert WTP reliability is “fair to moderate” based on a complete combinatorial test of means. Schaafsma et al (2014) report 57 percent choice congruency for land use changes after a year and “very good agreement” for WTP based on overlapping confidence intervals but mean WTP varied by minus 527 to plus 160 percent for some attributes. Lienhoop and Volker (2016) found that WTP for German forests did not vary significantly after a delay of one week. Czajkowski, Barczak, Budziński, Giergiczny, & Hanley (2016) report that WTP parameters for public forest management were significantly different after a 6 month delay, but that means were “relatively” stable. In contrast, Lew & Wallmo (2017) found no significant change in WTP for endangered species after 17 months. To summarise, stability of stated WTP for the environment appears to be the exception rather than the norm. It is apparent that utility maximisation theory provides only limited insight into these findings.

1.2 Constructed versus discovered preferences

There are two perspectives in behavioural decision research that can provide insight into apparent preference instability: discovered versus constructed preferences. The discovered preference hypothesis (DPH) was proposed by Plott (1996), who stated that when people have to make decisions about an unfamiliar issue or in an unfamiliar environment, their initial responses may be impulsive. As they learn about the decision environment (institutional learning) and their own attitudes (value learning), their decisions begin to exhibit less randomness and greater rationality. Preference discovery requires repetition, feedback on consequences and belief that those consequences are real. The requirement for feedback is important and some systematic biases have been reported to persist unless people experience a loss as a result of their choice (Braga & Starmer, 2005). However, it is problematic to provide feedback on consequences for environmental changes that may take years to eventuate. Lienhoop and Volker (2016) suggest that group discussion and reflection time may provide feedback and lead to more preference discovery than simple repetition, although they were not able to detect a statistically significant increase in preference adjustment. In our study about beach management preferences, DPH implies we might expect some institutional learning and a corresponding decrease in choice error in retests similar to that found in within-survey choice task repetition (Hess et al., 2012). “On the other hand we may not find any increase in value learning because our experiment did not include any mechanism by which respondents could gain feedback on the implications of their choices”.

The alternative constructed preference perspective is that preferences for the unfamiliar are often constructed, not merely revealed, when a decision is required (Gregory, Lichtenstein, & Slovic, 1993). This view rejects the usual presumption that stable and context-free preferences exist independently of the elicitation process and has been criticized for undermining the foundations of rational choice theory (Plott, 1996). However, consumers and voters make real-life decisions about unfamiliar products and issues regularly. Unfamiliarity, complex information, and public good character can cause instability in real-world choices as well as stated preferences (Carlsson, 2010) so a lack of pre-existing preferences does not necessarily invalidate SP methods. Similar to the ways by which authorities attempt to educate stakeholders during a policy consultation process; the role of the non-market valuation researcher is to ensure respondents have all the relevant information and make decisions with a high standard of reasoning (Gregory et al., 1993). When preferences are constructed rather than pre-existing they tend to be more strongly influenced by situational and framing effects such as presentation order (Krosnick & Alwin, 1987) or arbitrary anchors (Ariely, Loewenstein, & Prelec, 2003). Preferences may be constructed using a variety of simplifying strategies rather than expected utility maximisation. The result is that constructed preferences may be confined in scope (e.g. to a specific elicitation format) and transient – soon to be forgotten (Simon, Krawczyk, Bleicher, & Holyoak, 2008). The constructed perspective implies that preferences may not necessarily stabilise with repetition, especially if a time delay means that respondents don't remember their exact choices from the previous task.

The work presented in this paper is based on a fully-ranked choice experiment for erosion management options for beaches on the Coromandel Peninsula of New Zealand. We conduct not one but two identical re-tests each spaced three months apart. Having three points in time allows a more robust assessment of individual stability of stated WTP estimates in a manner that, as far as we are aware, no other study of environmental WTP has reported. Coastal landscapes are an important part of New Zealanders' identities (Collins & Kearns, 2010) and it is reasonable to assume respondents have pre-existing general preferences for coastal features and experience of beaches with the management options described. However, they have probably never been asked to make a specific trade-off between beach management and taxes so it is difficult to say whether the discovered or constructed viewpoint is likely to be more applicable. Our first research question is how stable is WTP in our specific context, and is this consistent with other test-retest studies? But the more interesting and unique question is does stability improve between the first and second re-test? If so, it would be consistent with the concept of learning and preference discovery. If not, the results would be more consistent with the transience of preferences constructed on the spot. We also investigate to what degree choice consistency can be explained by individual-specific factors. If preference stability could be predicted this could improve confidence in one-shot experiments where retest is not an option.

2. Method

2.1 Random utility models

Management options for Coromandel beaches may be thought of as a bundle of characteristics that affect the aesthetics and use of the beach. As per random utility theory (McFadden, 1974) we assume the probability of a consumer choosing their preferred future state of a beach is a function of deterministic and random or unobserved components of utility. Since the purpose of this study is to test for stability of WTP over time, we use a random utility model specified directly in "WTP-space" (Train & Weeks, 2005) such that the attribute parameters are interpretable as marginal WTP for each attribute.

Management options for Coromandel beaches may be thought of as a bundle of characteristics that affect the aesthetics and use of the beach. As per random utility theory (McFadden, 1974) we assume the probability of a consumer choosing their preferred future state of a beach is a function of deterministic and random or unobserved components of utility. Since the purpose of this study is to test for stability of WTP over time, we use a random utility model specified directly in "WTP-space" (Train & Weeks, 2005) such that the attribute parameters are interpretable as marginal WTP for each attribute. This is in contrast to the historically more common utility specification in "preference space" by which one first estimates preference parameters for attributes and cost (marginal utility of income) and then combines these to derive marginal WTP estimates. A model with utility specified in WTP-space is a more efficient estimator of WTP distributions and in random parameter models tends to produce spreads of marginal WTPs that are more plausible (Scarpa, Thiene, & Train, 2008). WTP-space models have previously been applied to outdoor recreation (e.g. in mountains by Scarpa et al., 2008 and in public forests by Czajkowski et al., 2016), as well as in other nonmarket valuation fields (e.g. in food choice by Balcombe et al. 2009 and in energy Scarpa and Willis 2010).

In this study we obtained full rankings of six alternatives in each choice card. The choice probabilities are modelled using the standard exploded logit model (Lancsar & Louviere, 2008). The utility in WTP space that person *n* obtains from the alternative state *j* and measured in time period *t* is specified as follows:

$$U_{njt} = \lambda_{nt}(ASC_j + \omega_{nt}'\mathbf{x}_j - p_j) + \varepsilon_{njt} \quad (1)$$

Where ASC is an alternative-specific constant for position on the choice card, $\mathbf{x}_j$ denotes the attribute levels of the non-price scenario, p_j is price, ε_{njt} is an i.i.d. extreme value type 1 error term, *n* are individual respondents and *j* are the alternatives. ω_{nt} is a vector of marginal WTP parameters specific to each individual *n* and assumed to be normally-distributed. λ_{nt} is a mixture of scale and cost coefficient with an assumed log-normal distribution to ensure the expected positive sign. Any unobserved variation in scale is also captured by this parameter. If we re-write indirect utility asV_{nit} (β_{nt}), with denoting the vector of random coefficients in equation 1, then the unconditional probability of person *n* set of choices in her sequence of *k* ranking over *t* repetitions is therefore the integral of the product of standard logit formulas over all values ofβ_{nt}:

$$P(i,t) = \int_{\beta} \prod_{t} \prod_{k} L_k(V_{nit}(\beta_{nt}))\varphi(\beta | b_{nt}, W_{nt}) d\beta \quad (2)$$

Where $\varphi(\beta | b_{nt}, W_{nt})$ is, in our case, normal densities with mean b_{nt} and var-covariance W_{nt}. This is known as a panel rank-exploded mixed logit specification and allows for taste variation across individuals, unrestricted substitution patterns and correlations in unobserved components across the choices by the same respondent (Train, 2002).

2.2 Re-test selection bias

If the decision to participate in the re-test is not independent of preference stability[1] then there is potential for selection bias in the results. As per a standard sample selection model (Winship & Mare, 1992) we specify that continuous latent variables Y_{1n}^{*} and Y_{2n}^{*} affects whether the choices of individual *n* are observed in retest 1 and 2. We fit binary logit models such that

$$Y_1 = 1 \text{ if } Y_{1n}^{*} > 0 \tag{3}$$

$$Y_1 = 0 \text{ if } Y_{1n}^{*} \leq 0 \tag{4}$$

to test whether there is a strong relationship between demographic variables and re-test participation. For respondents who complete the first retest we also test whether choice consistency is a significant explanatory variable for participation in the second retest.

2.3 Tests of stability

We test reliability of a DCE at three levels: (*i*) the proportion of identical choices, (*ii*) equality of the utility function and (*iii*) equality of marginal willingness-to-pay (WTP) for attributes, which is a less restrictive test of the equality of utility function.

2.3.1 Choice congruency

Comparison of choices is possible only when the same individuals are sampled in both the test and re-test. The measure of stability is the proportion of choice situations in which the same choice was made (congruency). Respondents may select the same alternative purely by chance so we correct this using Cohen's κ (Cohen, 1968), which acts as a correction factor for random matching:

$$\kappa = \frac{\mathrm{p_o - p_c}}{1 - \mathrm{p_c}} \tag{5}$$

where $\mathrm{p_o}$ is the observed probability and $\mathrm{p_c}$ is the probability that we would expect by chance.

2.3.1.1 Panel logit model for choice consistency

We estimate panel binary logistic regressions with random effects using R (2012) to explore the relationships between choice and individual characteristics and choice consistency in both retests.

[1] Selection bias may also affect average WTP in retests but this paper is concerned with consistency at an individual level

The dependent variable is one if the retest rank is the same as the rank in the first survey, otherwise it is zero. The set of binary outcomes can be written as:

$$P(Y_i = 1|\eta_i) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}, \qquad \eta_i = x_i\alpha + z_i u \tag{6}$$

where x_i are the fixed effects, α are the fixed effects parameters, z_i are the random effects and μ is the unobserved portion of heterogeneity. We include parameters for rank level, demographics and individual-specific variables that we expect to be related to emotional involvement or consumptive experience in the study area—travel distance, number of days visited in the previous year, and Coromandel holiday home ownership. Changes in individual circumstances might also cause people to adjust their preferences for beach recreation. We asked respondents if their household composition, income, labour force status or education level changed in each retest (wave 2 and 3) and included dummy variables for changed circumstances.

How restrictive choice congruency is becomes apparent when considering it would be difficult for someone to rank six alternatives in exactly the same order each time. It ignores the unobserved component of utility, which means that we should expect at least some degree of random error even if preferences are indeed stable. A change of one position is less inconsistent than a complete reversal of ranks so we also report the absolute difference in ranks for both waves. However, we do not estimate a linear regression for difference in rank because rank is an ordinal not a lineal variable and it would be incorrect to treat it as such. Another issue with using choice congruency as a measure of reliability is the risk that respondents may simply be remembering previous choices and selecting the same alternative rather than processing the information again, which would bias reliability upwards. Mørkbak and Olsen (2014) found no evidence of a memory effect on reliability after just two weeks so there is unlikely to be one in our case since the two waves were taken three months apart. However, we did not ask respondents whether they remembered their previous choices, so we are unable to specifically test this instance, which we expect to be quite remote.

2.3.2 Stability of parameter estimates

Testing for equality of the random utility function allows for random error in responses. We include only for respondents who completed all three waves so that sample differences are not confounded with stability measures. We follow the LR procedure detailed by Swait and Louviere (1993) in which the data from test and retest is stacked and a pooled model is estimated. The likelihood ratio (LR) test statistic is calculated as follows:

$$LR = -2\left(LL_{pooled} - \left(LL_1 + LL_2\right)\right) \tag{7}$$

where $LL_1$is the final log-likelihood of the model for the first test, LL_2 is for a retest, and LL_{pooled} is for the pooled model. The LR statistic has a Chi-square distribution with degrees of freedom equal to the number of parameters in the utility function. The LR is an asymptotic test of global goodness of fit and it tells us whether the variables with restricted (to be equal across waves) coefficients explain the same amount of variance before or after the restriction (Brouwer, 2006; Brouwer & Spaninks, 1999). If LR statistic does not exceed the five percent critical value we can conclude that the models for test and retest are sufficiently similar. The less restrictive test LR involves including explicit scale parameters in the pooled model to allow for differences in relative scale across waves. In a WTP-space model the scale parameter (λ) is in fact a combination of scale and the marginal utility of money. If the additional parameter is significant it is impossible to know whether one or both are different across waves, but a difference in λ does not affect WTP.

We also use Wald tests of joint asymptotic parameter equality between each pairwise combination of waves. The Wald test statistic, *W* is:

$$W = \left(\hat{b}_1 - \hat{b}_2\right)'\widehat{V}_1^{\,-1}\left(\hat{b}_1 - \hat{b}_2\right) \tag{8}$$

where $\hat{b}_1$ and $\hat{b}_2$ are the vectors of parameter estimates from models one and two and $\widehat{V}_1$ is the variance-covariance matrix of model one. Similar to the LR test, this statistic also has a sampling distribution that is asymptotically Chi-square distributed with degrees of freedom equal to the number of restricted parameters.

2.3.3 Stability of WTP

The joint tests do not identify which utility coefficients vary significantly from the restricted and unrestricted specification so we also perform equality tests on each WTP parameter. We use the variance-covariance matrix at convergence and Monte Carlo simulation (Krinsky & Robb, 1986) to approximate the asymptotic sampling distribution of WTP and use the non-parametric Mann-Whitney U test[2] (Brouwer & Spaninks, 1999) to test for equality of WTP means between each wave. We also examine the distributions of WTP using the Kolmogorov-Smirnov two-sample test because

[2] The Mann-Whitney U test (also known as Wilcoxon rank-sum) involves ranking the pooled WTP and then adding up the ranks for test and re-test datasets. The statistic U is given by:

$$U_1 = R_1 - \frac{n_1(n_2 + 1)}{2}$$

where $R_1$is the smaller sum of ranks of the two samples and n_1 and n_2 the sample sizes. For large samples U is approximately normally distributed with mean $\frac{n_1 n_2}{2}$ and standard deviation $\sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}}$.

this is a more restrictive null hypothesis than equality of means (Brouwer & Spaninks, 1999). The K-S test statistic *D* is sensitive to differences in both location and shape, is not reliant on normality and is based on the maximum absolute difference between the two cumulative distribution functions $S_{N_1}(x)$ and $S_{N_2}(x)$:

$$D = \max_{-\infty < x < \infty} \left| S_{N_1}(x) - S_{N_2}(x) \right| \tag{9}$$

The null hypothesis of equal distributions is rejected at level α if:

$$D_{mn} > c(\alpha)\sqrt{\frac{m+n}{mn}} \tag{10}$$

where *m* and *n* are the sample sizes and $c(0.05)$ equals 1.36 for sufficiently large samples.

3. Study design

3.1 The survey instrument

The data were collected in a web-based survey developed to gather information about preferences for beach management among domestic visitors to the Coromandel peninsula, New Zealand. The Coromandel is a steep and hilly peninsula that lies across the Hauraki Gulf from Auckland city. The peninsula is sparsely populated but is a popular holiday destination for residents of the nearby urban areas of Auckland and Hamilton, and to a lesser extent, international tourists. There are many beaches with high scenic and recreational appeal. Since the 1950s there has been considerable development pressure for holiday accommodation and some of the older developed areas are now at risk from coastal erosion. The primary purpose of the survey was to estimate the effect of different erosion management and headland development options on non-market value.

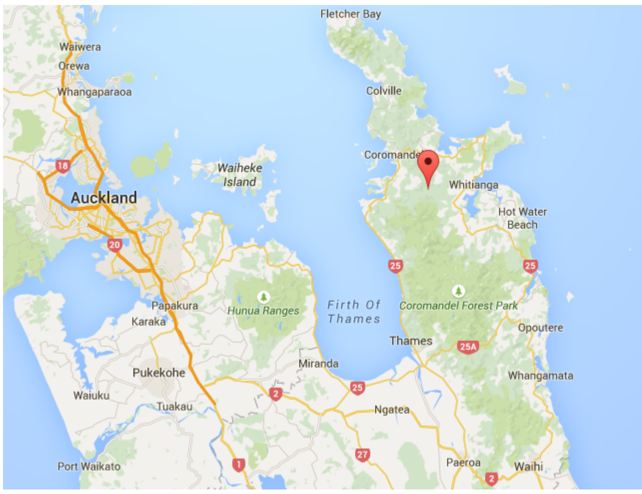

Figure 1 - Map showing the location of Coromandel Peninsula relative to Auckland (Source: Google Maps)

The survey included questions about their previous and planned beach visits, location of residence, environmental attitudes, socio-economic variables, and the choice experiment questions. Respondents were selected from a pre-recruited panel of New Zealand residents provided by a market research company and a smaller, self-selected sample from online advertisements on Facebook and Google. To qualify for the survey, respondent had to have visited the peninsula in the previous twelve months. Data collection was conducted in three separate waves in October 2013, January 2014 and April 2014 so as to gather additional information about recent beach trips and preference stability.

3.2 Experimental design

The choice experiment design was relatively simple with only three attributes—erosion protection, headland and cost—because virtual 3D models had to be created for each combination of attribute levels. Respondents were randomly assigned to a treatment group—who received videos, static images and text for the scenarios—and a control group who saw only static images and text. The video presentation format and impact is discussed in more detail in Matthews, Scarpa, & Marsh (2017). Table 1 shows the attribute levels and descriptions.

Table 1- Attributes and levels used in the choice experiment

Attribute	Description	Levels
Erosion protection	The beach is *x* km long and *y* km of this has properties at risk from erosion and high waves during storms. The options are to do nothing, remove the front row of properties and restore the nature dune system or build a seawall.	None Restored dune Sea wall
Headland	The headland is currently undeveloped and covered with native bush. If development is allowed then houses will be visible in future	No development Development allowed
Household taxes	Protection of the headland and foreshore require public funding so some of these options will increase your annual rates or taxes by the amount shown	$10 increments from $0 to $100

Respondents were given descriptions for three similar beaches of varying lengths with the current condition being no erosion protection and an undeveloped headland. Each choice card presented the respondent with six alternatives in random order so that every combination of headland and erosion protection appeared. This layout was tested with participants of a focus group who strongly preferred this to the alternative design of pair-wise alternatives where their preferred combination might not appear, even though it made their choice more complex. We generated a Bayesian-efficient design (Scarpa, Campbell, & Hutchinson, 2007) by swapping and cycling the cost attribute to minimise the average D-error across the distribution of prior values obtained from a focus group. The choice cards show thumbnail images of the attributes and a play button to play a video tour of the beach in a pop-up window. A sample choice card is provided in appendix 5.1. When survey respondents selected their preferred alternative it disappeared and they were asked to select the next preferred and so on until all six alternatives were ranked. We use an exploded logit format (Lancsar & Louviere, 2008) to model the ranks as repeated choices from sets with a decreasing number of alternatives. Respondents completed one choice card for each of the four beaches and one was selected at random to be used in the re-tests.

The choice questions were followed by a “stated certainty” question (Beck, Rose, & Hensher, 2013) in which the respondent was asked if they were sure they would have the same preference in real life if their preferred scenario was implemented in policy with the associated real local tax increase. The response format was a five-point scale comprising “definitely not”, “probably not”, “maybe”, “probably” and “definitely”. Self-reported stated certainty measures have been found to be a function of several individual characteristics and tend to be inversely correlated with choice error (Beck et al., 2013).

4. Results

4.1 Descriptive statistics

The sample for the first survey comprised 1059 individuals. There was considerable attrition over the six month period and only 551 completed the second wave and 426 completed the third wave. The final sample of individuals who completed all three waves was 387. Attrition is a major problem in panel studies: a drop-out rate of around 50% after the first survey is typical (Fitzmaurice, Heath, & Clifford, 1996). Table 2 shows a selection of demographic variables for the samples. Respondents tended to be older and more highly educated than the general population. There are small differences in means across waves for several variables (female, school children, high income, holiday house, travel distance and video treatment).

Table 2 - Descriptive statistics for each survey wave

Measure	Completed Wave 1	Completed Wave 2	Completed Wave 3	Completed all waves
Count of respondents	1059	551	426	387
Age (in years)	43	44	44	44
Degree	0.46	0.49	0.51	0.50
Female	0.59	0.63	0.63	0.63
Preschool children in household	0.17	0.16	0.17	0.17
School children in household	0.28	0.31	0.29	0.29
Annual household income < $50k	0.30	0.32	0.31	0.31
Annual household income > $100k	0.30	0.29	0.27	0.28
Holiday house owned by family	0.26	0.24	0.21	0.21

Travel time to site (hours)	2.33	2.27	2.19	2.20
Number of peninsula visits	2.36	2.57	2.29	2.31
Video treatment	0.52	0.56	0.61	0.60
Certain of choices	0.38	0.53	0.50	0.47^
Uncertain of choices	0.20	0.19	0.13	0.23^

^as measured in wave 1

4.1.1 Re-test selection bias

We fit two binary logistic regression models for retest participation (Table 9 in the appendix) for completion of retest one and two. The models have low explanatory power with pseudo R-squares of around 0.03 but there are some statistically significant effects. Women, respondents with school-age children and people in the video treatment group were more likely to re-participate. For second retest participation we include a variable for choice congruency from the first retest and it is insignificant. The lack of the significance of this variable combined with the fact that there is no overlap between significant variables for retest participation and those explaining choice congruency (Table 10 in the appendix) implies consistency results are unlikely to be affected by selection bias.

4.1.2 Choice congruency

Under a third (29 percent) of alternatives in the second wave were ranked identically to the first wave. While this is lower than the 57-59 percent congruency reported by Schaafsma et al. (2014) and Liebe et al. (2012), there are six fully ranked alternatives on the choice cards in this study rather than a single choice between three alternatives as in the other studies. If respondents selected randomly we would expect only 1/6 (17 percent) rank congruency. After adjusting for chance we calculate a Cohen's κ of 15 percent. There is higher rank congruency for the first rank (42 percent) and last rank (34 percent) than in the middle ranks (22 to 27 percent). This is consistent with the finding that choice error is lower for the best and worst alternatives (Ben-Akiva, Morikawa, & Shiroishi, 1992). Congruency between waves one and three is slightly lower at 26 percent, while the average for waves two versus three is 28 percent. The cumulative frequency graph (Figure 2) shows that half of the observations differ by only one position in waves two and three. Randomly simulated choices resulted in a median difference of two ranks. The rank difference is marginally larger in wave three than wave two.

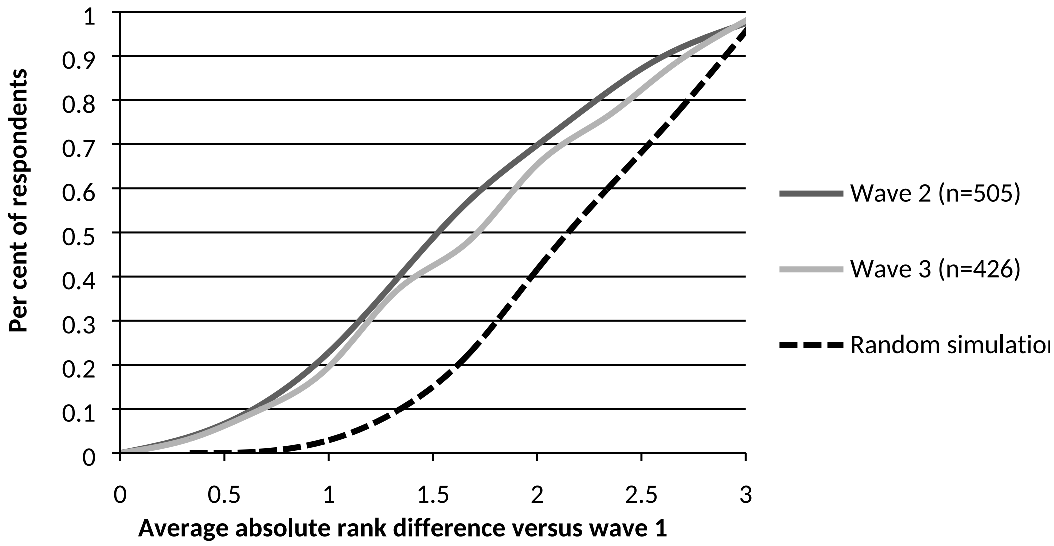

Figure 2 - CDF of absolute difference in ranks

Many respondents reported a change in household composition, income, labour force status or education level and these are reported in Table 3. Some people refused to answer a demographic question in one or more retests. The proportion of missing observations is high (up to 34 percent for household composition) which may attenuate any explanatory effect on choice congruency.

Table 3 - Changes to individual characteristics

	Retest 1		Retest 2	
Measure	Count	Missing	Count	Missing
Household change	99 (18%)	189 (34%)	99 (23%)	84 (20%)
Income increase	57 (10%)	136 (25%)	59 (14%)	77 (18%)
Income decrease	49 (9%)	136 (25%)	37 (9%)	77 (18%)
Labour force status change	71 (13%)	148 (27%)	68 (16%)	75 (18%)
Education level change	55 (10%)	144 (26%)	47 (11%)	60 (14%)

The logistic regressions for congruency (Table 10 in the appendix) have relatively poor overall model fit, indicating a large unobserved component to consistency. Education tends to be associated with lower within-survey choice error (Mazzotta & Opaluch, 1995) and also has a positive effect on choice consistency over time in our results. Ranks two to six have negative parameters so are less consistent than rank one. It is generally easier to choose the most preferred alternative (Ben-Akiva et al., 1992). Liebe et al. (2012) found choice consistency to be higher for the status quo alternative, but our status quo parameter is insignificant.

People with more experience with the good being valued tend to have better formed and more stable preferences (Brouwer, Dekker, Rolfe, & Windle, 2010; LaRiviere et al., 2014; McConnell et al., 1998). To test this hypothesis we include variables for ownership of a holiday house on the peninsula, travel distance and days spent visiting the peninsula in the previous year as measures of experience. We find that ownership of a holiday house is associated with higher choice congruency only for the first retest.

The video treatment effect on choice consistency is positive but insignificant. The video treatment is, however, positively correlated with stated certainty (people who answered “definitely” or “probably”) which is strongly positive and significant. This is in contrast to Mørkbak and Olsen (2014) who found a positive but insignificant relationship between stated certainty and retest consistency. We also test a variety of variables measuring a change in personal circumstances including income increase/decrease, gain/loss of employment, a change from single-person household to partnered to a family with children (and vice-versa), but find none of these to be significant predictors of choice congruency, similar to previous environmental test-retest choice experiments (Liebe et al., 2012; Schaafsma et al., 2014). Measurement error was perhaps too high to detect any effect even if it did exist.

4.2 Models and parameter equality

We estimated pooled and separate (for each wave) WTP-space random parameter logit models for respondents who completed all three waves using maximum simulated likelihood estimation in Biogeme (Bierlaire, 2003). Dune restoration, headland development, seawall and status quo alternative all have normally distributed random parameters while cost/scale parameter (λ) is log-normal. We also estimated similar separate and pooled models for the sub-sample of respondents who claimed to be certain (“definitely” or “probably”) of their choices in wave one.

Table 4 shows the values for the simulated log-likelihoods at convergence and the likelihood ratio (LR) test statistics. When including all respondents who completed all three waves, the LR test is significant at one percent even when allowing for scale/price coefficient differences. This means that the preference structure is significantly different across waves, not an uncommon finding in time-delayed test-retest surveys (Liebe et al., 2012; Schaafsma et al., 2014). However, using only “certain” respondents, the LR test statistic B is insignificant. This means that “certain” respondents did not significantly alter their preferences after allowing for variation in scale or marginal utility of money.

Table 4 - Pooled and separate model likelihood ratio tests for respondents who completed all 3 waves

Sample	LL Separate Models	Pooled A ($\lambda1 = \lambda2 = \lambda3$)	Pooled B ($\lambda1 \neq \lambda2 \neq \lambda3$)	LR test A	LR test B	H0: $\beta1=\beta2=\beta3$ Rejected?
All respondents	-6836	-7063	-6896	454.94***	120.55***	Yes
"Certain" respondents	-3130	-3195	-3143	129.32***	25.98	No

In Table 5 we show the results of Wald tests of joint parameter equality between each pair of waves. The tests reject joint parameter equality even for “certain” respondents. However, if we consider only parameter means and not the random parameter standard deviations, the tests are insignificant. This implies means but not variances are stable for “certain” respondents.

Table 5 – Pairwise Wald tests for respondents who completed all 3 waves

Sample	Parameters	Wave 1 vs 2	Wave 1 vs 3	Wave 2 vs 3
All respondents	All parameters	149.17***	691.85***	1084.47***
	Means only	17.44***	202.40***	96.60***
"Certain" respondents	All parameters	92.81***	77.37***	131.56***
	Means only	10.98	7.38	8.94

Table 6 reports the parameter estimates and their significance level for the separate waves, pooled model A with equal scale, pooled model B with unrestricted scale, and pooled model C with “certain” respondents only and unrestricted scale. It is encouraging that almost all parameters have stable signs and similar order of magnitude. The exception is the status quo coefficient estimate, which has a mean insignificantly different from zero in most cases, but often significant standard deviations. This simply suggests a large variation of the status-quo effect around zero across respondents and it is plausible. The alternative specific constants to control for position are significant in all models and do not decrease in significance in waves two or three. There is an enduring left-right bias that repetition does not erode, which is well documented in ranked and other choice data (Campbell & Erdem, 2015; Scarpa, Notaro, Louviere, & Raffaelli, 2011). The mean for dune restoration and seawalls are positive and headland development is negative, although the random parameter standard deviations are wide enough that a large chunk of the distributions are on the far side of zero. We expected significant heterogeneity in taste over attributes because people have different attitudes towards erosion protection and this is reflected in the significance of the random parameters. The relative importance of the attributes does not vary across waves or for “certain” respondents. The mean for headland development is always the largest in absolute terms, and the mean for seawalls the lowest.

Table 6 – Panel Random Parameter Logit models

Variable		Wave 1	Wave 2	Wave 3	Pooled A $\lambda1 = \lambda2 = \lambda3$	Pooled B $\lambda1 \neq \lambda2 \neq \lambda3$	Pooled C "Certain" $\lambda1 \neq \lambda2 \neq \lambda3$
Position 1 (left-most)		63.20*	19.00*	30.50***	57.80***	26.40**	59.30**
Position 2		72.60*	21.40*	46.20***	61.10***	26.40**	56.70**
Position 3		54.30*	18.60	43.00**	50.60**	22.90**	39.30*
Position 4		43.50	21.90*	36.50*	48.20**	22.90**	46.40*
Position 5		1.80	15.80	8.50*	26.30*	15.30	48.90*
Ln(λ)	μ	-4.73***	-4.26***	-5.32***	-4.97***	-8.19***	-6.48***
	σ	0.49***	1.76***	2.46***	0.48***	0.94***	-0.51***
Restored dune	μ	87.80**	32.00***	61.00***	55.30***	34.90***	56.50**
	σ	117.00**	72.30***	95.50***	63.10***	83.60***	81.90**
Headland development	μ	-84.20**	-76.40***	-155.00***	-88.70***	-85.80***	-84.80**
	σ	211.00**	95.10***	207.00***	81.40***	108.00***	95.90***
Seawall	μ	49.90*	10.40*	13.85*	32.40***	19.10***	41.60**
	σ	-204.00**	-78.80***	-148.00***	-84.10***	-72.60***	-116.00**
Status quo	μ	3.50	9.36	6.33	7.72	5.94	5.55
	σ	51.20	-11.60**	-80.90***	-13.90	-21.70**	51.40**
Scale parameter wave 2						3.85***	2.25***
Scale parameter wave 3						3.87***	1.94***
Number of observations		1960	1965	1965	5890	5890	2685
Number of individuals		387	387	387	387	387	180
Log-likelihood		-2295	-2304	-2299	-7141	-6976	-3143
Pseudo-R2		0.110	0.109	0.111	0.079	0.100	0.110
Bayesian information criteria		4704	4721	4711	14411	14100	6421

Significant at 10%, ** significant at 5%, *** significant at 1%

4.3 Willingness to pay

We present the results of the marginal WTP simulations as box plots in Figure 3 and in tabular format in Table 7. The confidence intervals all overlap but the means are visibly different. WTP variance is higher in the sub-sample of certain respondents, perhaps due to the difficulty of achieving statistical precision in a smaller sample size (180 versus 387 individuals in the full sample).

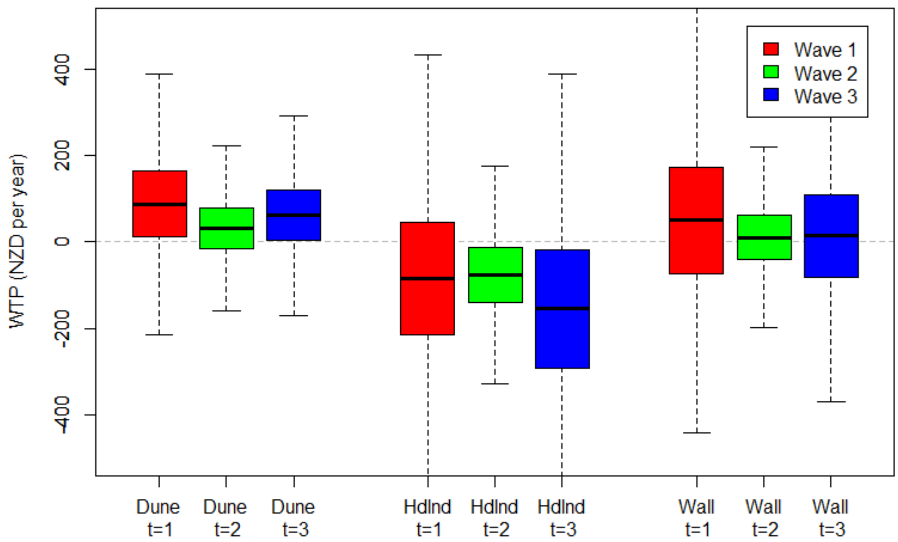

Figure 3 - Boxplot of WTP for respondents who completed all 3 waves

Table 7 - Mean WTP and confidence intervals for individuals who completed all 3 waves

	All respondents		"Certain" respondents	
	Mean	90% C.I.	Mean	90% C.I.
Wave 1				
Dune restoration	88	(-132,307)	70	(-265,404)
Headland developed	-84	(-467,299)	-88	(-473,295)
Seawall	50	(-317,417)	78	(-1024,1185)
Wave 2				
Dune restoration	32	(-89,153)	83	(-162,329)
Headland developed	-76	(-238,85)	-149	(-551,254)
Seawall	10	(-121,142)	47	(-320,411)
Wave 3				
Dune restoration	61	(-110,232)	59	(-122,241)
Headland developed	-155	(-503,194)	-96	(-310,119)
Seawall	14	(-235,263)	16	(-166,197)

Table 8 shows the results of the formal tests for mean and distribution equality as outlined in section 2. The Mann-Whitney U test is significant at five percent in seven out of nine cases and the

Kolmogorov-Smirnov test in every case. It follows that distributions of marginal WTPs are significantly different. For “certain” respondents we find some significant differences in mean WTP for headland development and in the variance of WTP for seawalls but WTP is otherwise stable.

Table 8 - Tests of equality of WTP means and distributions[3]

Attribute	Wave comparison	All respondents U-test (Z score)	All respondents K-S (D score)	"Certain" respondents U-test (Z score)	"Certain" respondents K-S (D score)
Dune restoration	1 vs 2	-6.20***	0.30***	-0.33	0.12
	1 vs 3	-1.98**	0.16***	-0.26	0.15*
	2 vs 3	-3.52***	0.19***	-1.00	0.14
Headland developed	1 vs 2	-0.12	0.19***	-1.58*	0.18*
	1 vs 3	-2.96***	0.17***	-0.09	0.14*
	2 vs 3	-4.96***	0.33***	-1.64*	0.22***
Seawall	1 vs 2	-2.33***	0.27***	-0.43	0.25***
	1 vs 3	-1.80**	0.14***	-0.77	0.35***
	2 vs 3	0.01	0.17***	-0.99	0.18**

Significant at 10%, ** significant at 5%, *** significant at 1%

4.4 Discussion and conclusion

This paper presents a study on temporal stability of WTP for beach development management. The study contributes to the limited research on temporal reliability in non-market valuation of environmental goods and has the unique feature of reporting on not only one but two retests and fully ranked choice cards. We find there is sufficient evidence to reject equality of joint and individual parameters in the WTP-space models in different time periods. Choice congruency is significantly higher than would be expected by chance alone, but there was little difference in congruency between waves one and two (29 percent), one and three (26 percent) and two and three (28 percent). Stability did not improve with the additional re-test, nor did left-right bias diminish. This implies that the tasks either lacked sufficient feedback to stimulate preference discovery, or that WTP was constructed on the spot as per the constructed preferences point of view. What we find to remain consistent is the relative importance of the attributes. The negative perception of headland development outweighed values for seawalls or dune restoration.

The implication for policy decision-makers is to be particularly cautious of stated preference values for goods that require complex and unfamiliar trade-offs, such as environmental quality. If values are to be used in a cost-benefit analysis we should focus on the order of magnitude of the values and the relative importance of the attributes. If the difference between cost and benefit is small, a high margin of error around the benefits will make it difficult to justify a decision.

[3] The number of draws used equals the comparison sample size in each case

On an encouraging note, we find there is a subset of respondents who exhibit more stable preferences. These respondents rated highly on scores of self-reported stated certainty. The use of certainty scores to measure a respondent's confidence in his or her choices originated from research on hypothetical bias (Beck et al., 2013) but our results suggest it may also be useful for predicting stability of preferences. Further research will be required to find out if this result is generalizable. Stated preference practitioners need to design experiments that maximise the likelihood of eliciting well-formed preferences (see Payne et al 1999 for a review of common faults in preference construction). Providing opportunity for deliberation might be useful (Lienhoop & Volker, 2016). Alternatively, researchers could attempt to make the consequences seem more real – for example, by providing virtual reality representation of the chosen scenario or personalised hypothetical rates invoices showing the cost. Certainty scaling questions could be used as a measure of relative success in this endeavour.

4.5 References

Ariely, D., Loewenstein, G., & Prelec, D. (2003). "Coherent Arbitrariness": Stable Demand Curves Without Stable Preferences. *The Quarterly Journal of Economics*, *118*(1), 73–106.

Beck, M. J., Rose, J. M., & Hensher, D. A. (2013). Consistently inconsistent: The role of certainty, acceptability and scale in choice. *Transportation Research. Part E, Logistics & Transportation Review*, *56*, 81. https://doi.org/10.1016/j.tre.2013.05.001

Ben-Akiva, M., Morikawa, T., & Shiroishi, F. (1992). Analysis of the reliability of preference ranking data. *Journal of Business Research*, *24*, 149–164. https://doi.org/10.1016/0148-2963(92)90058-j

Bierlaire, M. (2003). BIOGEME: a free package for the estimation of discrete choice models.

Bliem, M., Getzner, M., & Rodiga-Laßnig, P. (2012). Temporal stability of individual preferences for river restoration in Austria using a choice experiment. *Journal of Environmental Management*, *103*, 65–73. https://doi.org/10.1016/j.jenvman.2012.02.029

Braga, J., & Starmer, C. (2005). Preference anomalies, preference elicitation and the discovered preference hypothesis. *Environmental and Resource Economics*, *32*(1), 55–89.

Brouwer, R. (2006). Do stated preference methods stand the test of time? A test of the stability of contingent values and models for health risks when facing an extreme event. *Ecological Economics*, *60*(2), 399–406. https://doi.org/10.1016/j.ecolecon.2006.04.001

Brouwer, R., Dekker, T., Rolfe, J., & Windle, J. (2010). Choice Certainty and Consistency in Repeated Choice Experiments. *Environmental and Resource Economics*, *46*(1), 93–109. https://doi.org/10.1007/s10640-009-9337-x

Brouwer, R., & Spaninks, F. (1999). The Validity of Environmental Benefits Transfer: Further Empirical Testing. *Environmental and Resource Economics*, *14*(1), 95–117. https://doi.org/10.1023/A:1008377604893

Bryan, S., Gold, L., Sheldon, R., & Buxton, M. (2000). Preference measurement using conjoint methods: an empirical investigation of reliability. *Health Economics*, *9*(5), 385–395.

Campbell, D., & Erdem, S. (2015). Position bias in best-worst scaling surveys: a case study on trust in institutions. *American Journal of Agricultural Economics*, *97*(2), 526–545. https://doi.org/10.1093/ajae/aau112

Carlsson, F. (2010). Design of Stated Preference Surveys: Is There More to Learn from Behavioral Economics? *Environmental and Resource Economics*, *46*(2), 167–177. https://doi.org/10.1007/s10640-010-9359-4

Carlsson, F., Mørkbak, M. R., & Olsen, S. B. (2012). The first time is the hardest: A test of ordering effects in choice experiments. *Journal of Choice Modelling*, *5*(2), 19–37.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213.

Collins, D., & Kearns, R. (2010). "It's a gestalt experience": Landscape values and development pressure in Hawke's Bay, New Zealand. *Geoforum*, *41*(3), 435–446. https://doi.org/10.1016/j.geoforum.2009.11.010

Czajkowski, M., Barczak, A., Budziński, W., Giergiczny, M., & Hanley, N. (2016). Preference and WTP stability for public forest management. *Integrating Ecosystem Service Concepts into Valuation and Management Decisions*, *71*, 11–22. https://doi.org/10.1016/j.forpol.2016.06.027

Fitzmaurice, G. M., Heath, A. F., & Clifford, P. (1996). Logistic Regression Models for Binary Panel Data with Attrition. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *159*(2), 249–263. https://doi.org/10.2307/2983172

Gregory, R., Lichtenstein, S., & Slovic, P. (1993). Valuing environmental resources: a constructive approach. *Journal of Risk and Uncertainty*, *7*(2), 177–197.

Hess, S., Hensher, D. A., & Daly, A. (2012). Not bored yet – Revisiting respondent fatigue in stated choice experiments. *Transportation Research Part A: Policy and Practice*, *46*(3), 626–644. https://doi.org/10.1016/j.tra.2011.11.008

Krinsky, I., & Robb, A. L. (1986). On Approximating the Statistical Properties of Elasticities. *The Review of Economics and Statistics*, *68*, 715–719.

Krosnick, J. A., & Alwin, D. F. (1987). An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement. *The Public Opinion Quarterly*, *51*(2), 201–219. https://doi.org/10.1086/269029

Lancsar, E., & Louviere, J. (2008). Estimating individual level discrete choice models and welfare measures using best-worst choice experiments and sequential best-worst MNL. *University of Technology, Centre for the Study of Choice (Censoc)*, 8–4.

LaRiviere, J., Czajkowski, M., Hanley, N., Aanesen, M., Falk-Petersen, J., & Tinch, D. (2014). The value of familiarity: Effects of knowledge and objective signals on willingness to pay for a public good. *Journal of Environmental Economics and Management*, *68*(2), 376–389. https://doi.org/10.1016/j.jeem.2014.07.004

Lew, D. K., & Wallmo, K. (2017). Temporal stability of stated preferences for endangered species protection from choice experiments. *Ecological Economics*, *131*, 87–97.

Liebe, U., Meyerhoff, J., & Hartje, V. (2012). Test-Retest Reliability of Choice Experiments in Environmental Valuation. *Environmental and Resource Economics*, *53*(3), 389. https://doi.org/10.1007/s10640-012-9567-1

Lienhoop, N., & Volker, M. (2016). Preference Refinement in Deliberative Choice Experiments for Ecosystem Service Valuation. *Land Economics*, *92*(3), 555–577. https://doi.org/10.3368/le.92.3.555

Matthews, Y., Scarpa, R., & Marsh, D. (2017). Using virtual environments to improve the realism of choice experiments: A case study about coastal erosion management. *Journal of Environmental Economics and Management*, *81*, 193–208. https://doi.org/10.1016/j.jeem.2016.08.001

Mazzotta, M. J., & Opaluch, J. J. (1995). Decision making when choices are complex: a test of Heiner's hypothesis. *Land Economics*, *71*(4), 500–515.

McConnell, K. E., Strand, I. E., & Valdés, S. (1998). Testing Temporal Reliability and Carry-over Effect: The Role of Correlated Responses in Test-retest Reliability Studies. *Environmental and Resource Economics*, *12*(3), 357–374. https://doi.org/10.1023/A:1008264922331

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics*. New York: Academic Press.

Mørkbak, M. R., & Olsen, S. B. (2014). A within-sample investigation of test–retest reliability in choice experiment surveys with real economic incentives. *Australian Journal of Agricultural and Resource Economics*.

Plott, C. R. (1996). Rational Individual Behavior in Markets and Social Choice Processes: the Discovered Preference Hypothesis. In *The Rational Foundations of Economic Behaviour, ed. K. J. Arrow, E. Colombatto, M. Perleman, & C. Schmidt* (pp. 225–250). London: Macmillan. Retrieved from https://books.google.co.nz/books?id=wsO7QgAACAAJ

R Core Team. (2012). R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012.

Rigby, D., & Burton, M. (2011). Intertemporal choice consistency and the information sensitivity of welfare estimates in stated preference studies. Presented at the EAERE 18th annual conference, Rome.

Ryan, M., Netten, A., Skåtun, D., & Smith, P. (2006). Using discrete choice experiments to estimate a preference-based measure of outcome—an application to social care for older people. *Journal of Health Economics*, *25*(5), 927–944.

Scarpa, R., Campbell, D., & Hutchinson, W. G. (2007). Benefit estimates for landscape improvements: sequential Bayesian design and respondents' rationality in a choice experiment. *Land Economics*, *83*(4), 617–634.

Scarpa, R., Notaro, S., Louviere, J., & Raffaelli, R. (2011). Exploring Scale Effects of Best/Worst Rank Ordered Choice Data to Estimate Benefits of Tourism in Alpine Grazing Commons. *American Journal of Agricultural Economics*, *93*, 809–824. https://doi.org/10.1093/ajae/aaq174

Scarpa, R., Thiene, M., & Train, K. (2008). Utility in Willingness to Pay Space: A Tool to Address Confounding Random Scale Effects in Destination Choice to the Alps.

American Journal of Agricultural Economics, *90*, 994–1010. https://doi.org/10.1111/j.1467-8276.2008.01155.x

Schaafsma, M., Brouwer, R., Liekens, I., & De Nocker, L. (2014). Temporal stability of preferences and willingness to pay for natural areas in choice experiments: A test–retest. *Resource and Energy Economics*, *38*, 243–260.

Simon, D., Krawczyk, D. C., Bleicher, A., & Holyoak, K. J. (2008). The transience of constructed preferences. *Journal of Behavioral Decision Making*, *21*(1), 1–14.

Skjoldborg, U. S., Lauridsen, J., & Junker, P. (2009). Reliability of the discrete choice experiment at the input and output level in patients with rheumatoid arthritis. *Value in Health*, *12*(1), 153–158.

Swait, J., & Louviere, J. (1993). The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models. *Journal of Marketing Research*, *30*, 305–314.

Train, K. (2002). *Discrete choice methods with simulation*. Cambridge Univ Pr.

Train, K., & Weeks, M. (2005). Discrete choice models in preference space and willingness-to-pay space. In *R.Scarpa & A. Alberini, eds, "Applications of simulation methods in environmental and resource economics."* Springer Publisher, Dordrecht, The Netherlands.

Winship, C., & Mare, R. D. (1992). Models for Sample Selection Bias. *Annual Review of Sociology*, *18*, 327–350.

5. Appendix

5.1 Choice card

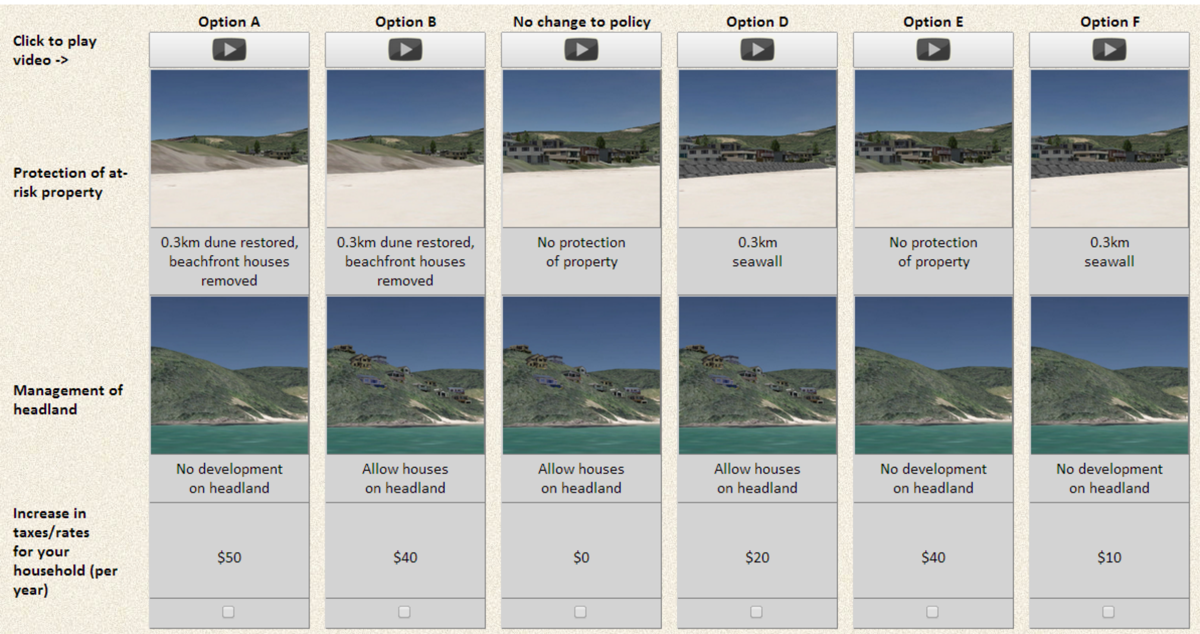

Figure 4 - Example beach choice card

Table 9 – Binary logit model results for retest participation

Dependent variable = retest participation	1st retest		2nd retest | 1st retest	
Variable	Coefficient	Z score	Coefficient	Z score
Constant	-0.335	-0.61	-0.735	-1.25
Age (in years)	-0.0244	-0.96	-0.0029	-0.11
Age squared	0.0004	1.46	0.0002	0.56
Degree	0.2363*	1.73	0.3540**	2.49
Female	0.2981**	2.18	0.2175	1.52
Preschool children in household	-0.0366	-0.20	0.1447	0.77
School children in household	0.4397***	2.91	0.1043	0.66
Annual household income< $100k	0.0255	0.15	-0.1865	-1.05
Annual household income > $100k	-0.0113	-0.07	-0.3028*	-1.67
Bach owned by family	-0.1747	-1.15	-0.3999**	-2.45
Travel time to site (hours)	-0.0451	-1.13	-0.1182**	-2.17
Peninsula visits duration (days)	-0.0157	-1.18	-0.0374**	-2.37
Video treatment	0.4175***	2.94	0.3925***	2.61
Certain of choice	-0.0929	-0.64	0.1002	0.66
Choice congruency 1st retest				
Number of individuals		1059		505
Log-likelihood		-1423		-651
Pseudo-R^2		0.029		0.036
Bayesian information criteria		2953		1389

Table 10 - Logistic regression for rank congruency

Dependent variable = 1 if ranks are the same as first wave, otherwise = 0				
	Wave 2		Wave 3	
Variable	Coefficient	Z -value	Coefficient	Z -value
Intercept	-0.982***	-3.07	-1.575***	-3.06
Rank 2	-0.778***	-5.25	-0.785***	-3.86
Rank 3	-1.078***	-7.01	-0.581***	-2.93
Rank 4	-0.858***	-5.72	-0.676***	-3.38
Rank 5	-1.128***	-7.27	-0.463**	-2.37
Rank 6	-0.413***	-2.85	-0.241	-1.26
Status quo alternative	0.090	0.75	0.118	0.76
Age (years)	-0.006	-1.31	0.010	1.63
Degree	0.300**	2.37	0.151	0.86
Female	-0.056	-0.43	0.327*	1.78
Preschool children in household	-0.157	-0.92	-0.012	-0.05
School children in household	0.046	0.35	0.126	0.69
Annual household income < $50k	-0.064	-0.41	-0.236	-1.09
Annual household income > $100k	-0.174	-1.11	-0.197	-0.93
Holiday home owned by family	0.295**	2.10	-0.036	-0.17

Travel time to site (hours)	0.014	0.27	-0.097	-0.71
Days visited peninsula	-0.002	-0.24	-0.015	-1.46
Video treatment	0.200*	1.68	0.171	1.39
Certain of choices	1.058***	8.12	0.450***	2.62
Change in income	-0.254*	-1.71	0.337	1.64
Change in labour force status	0.062	0.36	-0.340	-1.35
Change in household composition	-0.055	-0.36	0.347	1.57
Sigma (panel variance)	1.188***	10.92	1.344***	9.74
Number of individuals		551		426
Log-likelihood		-1659		-999
Pseudo-R2		0.087		0.062
Bayesian information criteria		3470		2142

- We repeat a choice experiment three times in six months with the same individuals
- Tests reject joint parameter equality and mean WTP equality
- Consistency does not improve in second retest
- Respondents with high self-reported certainty do have stable WTP