Interviewer effects in a discrete choice experiment implemented as a valuation workshop

Margrethe Aanesen[1]*, Erlend D. Sandorf*, Claire Armstrong*

Abstract

Human beings benefit from nature's ecosystem services. Many of these services have no price that can contribute to limit the consumption of them. One tool for deriving such prices is to derive Willingness-to-pay (WTP) estimates. Such estimates are often based on surveys where people are asked, directly or indirectly, about their WTP for benefitting from specific services provided by nature (stated preferences surveys). There are a few studies showing that personal characteristics of the interviewer in such surveys significantly affect the respondents' answers, whereas no research so far has been able to find significant interactions between personal characteristics of the interviewer and respondent. The present study reports from an experiment where we implemented identical valuation workshops, first with a middle aged female moderator, and next with a young male moderator. The results show significant differences in the responses across the two datasets, also when we correct for differences in sample characteristics. The differences in responses are further reinforced when we take into account personal characteristics of the respondents, but these effects are not statistically significant. Our results support earlier results that "non-good attributes", such as the characteristics of the interviewer, affect the stated preferences in f2f surveys. This may be an additional argument for conducting internet, or non-f2f surveys. The novelty of the experiment is that it takes place within a discrete choice experiment setting, whereas all previous comparable studies have used contingent valuation surveys.

JEL: Q51, Q57

Keywords: discrete choice experiment, valuation workshop, interviewer effects, interviewer-respondent interactions

[1] corresponding author: Margrethe.aanesen@uit.no

*University of Tromso – Arctic University of Norway, Faculty of Biosciences, Fisheries and Economics, P.O.Box 6050 Langnes, 9037 Tromso, Norway

1 Introduction and background

The use of stated preference (SP) techniques has become increasingly accepted and applied elements of valuation of non-market goods. However, due to the hypothetical character of these elicitation techniques, many are still strongly critical of the validity of SP survey results and thus the appropriateness of their use in cost-benefit assessments of public developments (see e.g. Spangenberg and Settele, 2010). In order to meet such objections it is necessary to further investigate the extent to which contextual issues may influence the way in which respondents formulate answers to questions asked in stated preference surveys.

While the different stated preference models are well grounded in neoclassical microeconomic theory, and theoretical concepts such as incentive compatibility and consequentiality can guide the formulation of the survey questionnaire, we also have to include other disciplines in order to understand how contextual issues influence the responses in SP surveys. One branch of such investigations is studies comparing different modes of administering SP surveys. Lindhjem and Navrud (2011) give a thorough review of results from comparisons of internet surveys and f2f or mail surveys, while Sandorf et al (2015) compare an internet survey with valuation workshops. Another branch of this literature concentrate on surveys with (personal) contact between interviewer and respondent, such as f2f surveys, and shows how the changes in personal characteristics of the interviewer, such as gender, race, ethnicity and language influence the respondents' answers in interview situations. Gong and Aadland (2011), Bateman and Mawby (2004), and Loureiro and Lotade (2005) report that responses to similar willingness-to-pay (WTP) questions vary significantly across different interviewers when varying specific characteristics of the interviewer. On the other hand, interactions between respondent personal characteristics and interviewer personal characteristics did not turn out significant (Gong and Aadland, 2011). This paper contributes to the current literature on interviewer effects, by showing that respondents in a discrete choice experiment implemented as a valuation workshop state significantly different WTP with different workshop moderators. Deriving this result we have corrected for differences in sample characteristics and in other exogenous factors of the interview situation as e.g. time and location.

In stated preference (SP) valuation surveys the mere presence of an interviewer, independent of personal characteristics, may affect respondents responses. Leggett et al. (2003) find that stated willingness-to-pay (WTP) in a contingent behavior (CV) survey is between 23-29% higher in f2f interviews compared to a self-administered survey, i.e. when the respondents sit in isolation when responding to the questionnaire. In both cases the respondents are asked exactly the same questions, and the reason for the higher WTP in the f2f-survey is explained by social desirability, i.e. that people provide answers that they believe the interviewer will like. Somewhat related is the study by List et al. (2004), which shows that respondents more frequently vote “yes” to a monetary contribution to the provision of a public good the more socially transparent the context is. For example are respondents in a peer group are on average 18% more likely to vote “yes” to support the public good compared to respondents who implement the survey in isolation and anonymously. Interestingly, List et al. (2004) show that survey design is more important to explain differences in response than whether the elicitation process is hypothetic or actual. If people believe that the outcome of the survey will have consequences for public policy, and they are interested in the issues under consideration, they are likely to answer according to their true preferences in hypothetical situations.

There are a few SP-surveys testing for personal characteristics of the interviewer. Bateman and Mawby (2004) let half of the sample be interviewed by a casually dressed moderator and the other half by a formally dressed moderator. The sample was further split so that one half got additional information. While there was no significant effect on WTP of additional information, the WTP rose significantly when the moderator was formally dressed. Loureiro and Lotade (2005) show that when interviewed in a supermarket, people are willing to pay a higher price-premium for eco-labelled or fair-trade products from African countries when interviewed by a person with an African-American look compared to when interviewed by a white American. Finally, Gong and Aadland (2011) find that respondents state higher WTP for recycling services when interviewed by white and female interviewers than by non-white and male interviewers. As far as we know, there are no tests of effects of interviewer characteristics on responses in discrete choice experiment (DCE) surveys, as all previous interviewer effect tests in SP studies have been implemented as part of CV surveys. One could argue that WTP-bias due to characteristics of the interviewer is more likely to occur in CV-surveys compared to in DCE-surveys because only in the former do we directly ask for WTP for a non-market good. For instance, the social desirability effect could be argued to be lower in a DCE-survey as we do not ask people directly to contribute monetarily. On the other hand, it is well-known that the presentation of the good to be valued is essential to the outcome of a valuation study, be it CV or DCE. If the valuation scenario is not convincing or there is lack of correspondence between the outcome of the survey and the realization of the scenario (consequentiality), survey results will often tend to be invalid (Kling et al., 2012). This is true for both CV and DCE surveys. Hence, the question of whether and how personal characteristics of the interviewer affect the respondents is of interest in both types of surveys.

Expanding upon the gender-effect issue there is also the question whether people respond differently to male and female interviewers depending on their own gender. As an example, Fuchs (2009) analyses interviewer gender - respondent gender interactions in a video-enhanced web-survey concerning sexual relations and sexually transmitted diseases. Whereas the interviewer gender – respondent gender interactions are not significant and go in same-gender as well as in opposite-gender directions, his results do indicate that gender-of-interviewer effects occur. A similar outcome is reported by Kane and Macaulie (1993) who report that both female and male respondents express more egalitarian gender-related attitudes or greater criticism of existing gender inequalities to female interviewers, which can easily be interpreted as a type of social desirability effect. They find that the effect of the interviewer's gender is significant in most attitudinal domains, but that the interaction between interviewer gender and respondent gender is not statistically significant. To our knowledge, the interaction between interviewer and respondent when it comes to personal characteristics and the consequences for the resulting stated preferences has not been analyzed in the context of DCE surveys before.

This paper presents the results of a DCE survey concerning the protection of cold-water-coral (CWC) implemented in a group setting; first by a middle aged female moderator and next by a young male moderator. CWC is an unfamiliar good for most people and most of the Norwegian population does not know of the existence of CWC off the Norwegian coast. Hence, it is likely to assume that the individual preferences of the respondents when it comes to protection of CWC are the same across the two samples. However, in addition to different interviewers, the first survey sampled respondents from 22 locations in Norway whereas the second only sampled 4 locations (a subset of the 22). Furthermore, there is a full year between data collections, which implies that there might be systematic differences in preferences between the two datasets. In section 2 we present data and methods. Section 3 presents the results, and section 4 discusses the results whereas section 5 concludes.

2 Data and method

2.1 Method

Working on two separately collected datasets, introduces a few issues which must be addressed in order to make comparisons of results.

1: Individual characteristics of the respondents may differ across the two datasets.

2: The scale of the utility might differ across the two datasets.

Whereas 1) can be controlled for by balancing the two datasets (Deheija and Wahba, 2002), 2) can be addressed by estimating the relative scale parameter (Adamowicz et al. 1994; Swait and Louviere 1993). Scale heterogeneity across the datasets implies that respondents in one dataset respond systematically different compared to respondents in the other dataset. The fact that the data in the two datasets has been collected under different conditions, e.g. at different time and by different interviewers, may lead to variation across respondents, which is heterogeneity that cannot be linked to measureable aspects of the respondents (Train 2009). Such variation would impact on all parameters in the same way, and materializes through differences in the scale parameter between the two datasets. There is of course also the possibility that scale differs across all individuals, and to isolate the part of the variation which is due to interviewer we formulate the scale parameter as a function of interviewer. In addition, we have to take into account individual preference heterogeneity, which is typically done by applying the random parameter (mixed) multinomial logit model MMNL. Combining random parameters and non-unit scale takes us to the generalized multinomial logit model (G-MNL). This model is sufficiently flexible to also test for possible interviewer – respondent interactions. Including "interviewer" as an explanatory variable in the G-MNL, excludes the possibility to let the relative scale also be interviewer-dependent, as this will "over-identify" the interviewer effect and yield confounded estimates for both (Fiebig et al., 2011, Hess and Rose, 2012). Hence, we identify the interviewer effect by letting the scale be interviewer-dependent, and let utility be a (linear) function of attributes of the good (coral protection) only.

2.2 Balancing the datasets by the use of Propensity Score Matching (PSM)

Differences in responses from two samples may be due to differences in respondent characteristics in the two samples. In order to secure comparable samples we applied the propensity score matching technique to balance the two datasets. This was necessary as we found significant difference in the individual characteristics across the two survey samples (see table 1a). Merging the responses from the two surveys would imply an unbalanced dataset, both in number of observations and in respondent characteristics. Working on an unbalanced dataset will give confounded results when analyzing eventual treatment effects (Austin, 2011). In order to achieve a balanced dataset, i.e. a dataset where treated and control respondents do not vary significantly w.r.t. personal characteristics, we applied the PMS technique (see e.g. Dahejia and Wahba, 2002, Austin, 2011). This worked as follows: First, we merged the two datasets. Based on the merged dataset we estimated (logistically) the probability for being interviewed by a male (treatment) explained by the personal characteristics of the respondents (12 covariates). Using the resulting coefficients for the covariates we calculated propensity scores (PS), i.e. the propensity to belong to the treatment group (interviewed by a male), for each of the respondents in the merged dataset. These propensity scores were finally used to match each of the treatment respondents with one respondent in the control group (interviewed by a woman). The matching was done based on similarity in PS.

We only matched respondents with equal PS score based on the two first 2-decimals in the PS, which resulted in a merged balanced dataset with 92 respondents from each of the male and female datasets. The remaining 14 respondents from the male dataset could not be matched by any respondent from the female dataset due to differences in PS-score.

2.3 The generalized multinomial logit model (G-MNL)

When the survey respondents are indexed *n*, the alternative *j*, and the choice situation *t*, the utility to individual *n* of choosing alternative *j* in situation *t* can be expressed by

$$U_{njt}^{*} = V_{njt} + e_{njt}^{*} \tag{1}$$

where V_{njt} is the explainable portion of the utility and e_{njt}^{*} is an idiosyncratic error expressing omitted variables which influence the utility of person *n* when choosing alternative *j* in situation *t*.

The explainable part of the utility is assumed to be a linear combination of attributes, denoted X_{njt}, each given a separate weight, which in the absence of individual preference heterogeneity is similar across respondents, β. Hence, we have

$$V_{njt} = \beta * X_{njt} \tag{2}$$

where X_{njt} is the vector of attributes of all alternatives j=1,…N.

In G-MNL the estimated parameters vary across individuals according to

$$\beta_n = \sigma_n \beta + [\gamma + \sigma_n(1-\gamma)]L\eta_n \tag{3}$$

where σ_n is the individual-specific scale of the idiosyncratic error term, γ is a scale parameter that controls how the variance of residual taste heterogeneity, $L\eta_n$, varies with scale.

In order to let σ_n vary across groups of individuals, i.e. all individuals interviewed by the male moderator, instead of across all individuals, we reformulate it as follows:

$$\sigma_n = exp(d * interv_n) \qquad (4)$$

where $interv_n$ is a dummy indicating whether the interviewer for individual *n* was male (=1) or female (=0). This implies that $\sigma_n = 1$ for respondents in the female dataset and $\sigma_n = exp^d$ for respondents in the male dataset, and thus that the relative scale parameter for the male dataset is (exp)^d .

For simplification, we set $\gamma = 1$, which implies that the residual taste heterogeneity is proportional to the individual scale.

When the interviewer effect is included as a dummy in the model, we get

$$V'_{njt} = \beta * X_{njt} + m_j I_{njt} \qquad (2')$$

where I is a dummy variable taking the value 1 when the interviewer is male and zero otherwise, and *m* is a parameter to be estimated. Note, that including both a dummy for interviewer and a relative scale parameter will imply "double counting" of the interviewer effect. Thus, *m* and *d* cannot be estimated simultaneously.

The unexplained portion of the utility, e^*_{njt}, is extreme value distributed (Gumbel) with variance $\sigma^2 \frac{\pi^2}{6}$. Here, σ^2 is the variance of the unobserved part of utility, and since this is irrelevant to respondent behavior it is usually assumed to equal one. This means that we assume all respondents to have the same variance of the unobserved factors. There is an inverse relationship between error variance and scale. A small error variance implies a large scale and consequently larger model parameters and a large error variance implies a small scale and smaller model parameters.
In our case, with data collected under two different circumstances, it is not unlikely that the unobserved factors have a variance that differ across respondents in the two datasets, implying that the variance of the unobserved factors is given by $\sigma_i^2 \frac{\pi^2}{6}$, i=F,M, for respondents in the female and male datasets respectively. The ratio of the variance is given by $(\frac{\sigma_M}{\sigma_F})^2$. Let $s = \sqrt{(\frac{\sigma M}{\sigma F})^2}$, then the choice probabilities are given by

$$P_{njt} = \frac{exp^{-b\prime Xnjt}}{\sum_{k=1}^{N} exp^{-b\prime Xnkt}} \tag{3}$$

for respondents in the female dataset and

$$P_{njt} = \frac{exp^{-(b/s)\prime Xnjt}}{\sum_{k=1}^{N} exp^{-(\frac{b}{s})\prime Xnkt}} \tag{4}$$

for respondents in the male dataset.

Applying maximum likelihood to estimate attribute weights, β, the Likelihood function is

$$LL = \sum_{n\in F} \log \prod_{t=1}^{T} \left[\frac{\exp(-\beta X_{njt}))}{\sum_k \exp(-\beta \quad X_{nkt}))}\right] \; + \sum_{n\in M} \log \prod_{t=1}^{T} \left[\frac{\exp(-(\frac{\beta}{s}) X_{njt}))}{\sum_k \exp(-(\frac{\beta}{s}) \quad X_{nkt}))}\right] \tag{5}$$

The relative scale parameter, s, is estimated together with the attribute coefficients, β.

2.4 Data

Cold water coral (CWC) is abundant off the Norwegian coast. Still, they have been impacted over many years mainly due to bottom trawling, but also sea bed operations as oil exploration pose a threat to them (Fosså et al., 2002). In order to inform Norwegian authorities on the Norwegian population's valuation of protecting CWC, we set up a discrete choice experiment (DCE) with two protection scenarios and the status quo (SQ). The attributes of the DC and the level they take are given in table 1.

Table 1 Attributes and attribute levels

Attribute	Size of protected area (km^2)	Protected area attractive for oil/gas and fisheries activities?	Protected area important as habitat for fish?	Additional costs of protection
SQ	2.445	Partly	Partly	0
Level 1	5.000	Attractive for the fisheries	Not Important	100
Level 2	10.000	Attractive for oil/gas activities	Important	200
Level 3		Attractive for both fisheries and oil/gas activities		500
Level 4		Neither attractive for fisheries nor for oil/gas activities		1000

During spring 2013 we implemented 24 group interviews in the form of valuation workshops in 22 Norwegian municipalities encompassing a total of 402 respondents. The group interviews were led by a female moderator aged 49, and a female assistant aged 35. Information yielded to the participants during the group interviews and the survey questionnaire is available from the authors upon request. During spring 2014 we implemented 6 similar group interviews (valuation workshops) in 4 Norwegian municipalities encompassing a total of 106 respondents. These group interviews were led by a male moderator aged 26, and the same female assistant as in the previous survey. The group process and survey materials were exactly the same. The same recruitment company recruited the participants to both surveys. The company was instructed to recruit a sample of the population in the municipality where the workshops took place, which was representative with respect to gender and age. These instructions were the same for each survey. Also, a written manuscript for the recruiter to read to inform potential respondents was identical for the two surveys. The company applied the same set of recruiters for both surveys. The characteristics of the respondents from each of the datasets are summarized in table A1 in the appendix.

We tested whether the respondents in the female and male datasets could be assumed to be randomly drawn units from one and the same population. For this purpose we applied Pearson's chi-squared test for comparing proportions from two samples (sex (proportions of males), ENGO-membership, participation in the labor force, working in the marine sector, living on the coast, and living in an urban area), and the Welch two-sample t-test to test for differences in personal characteristics measured as continuous variables or factors in the two samples (age, education, household size, personal income and household income. The reported p-value indicates the probability for the observed difference in the covariate to occur by chance alone. Hence, high p-values indicate no systematic difference in the distribution of personal characteristics in the two samples. On the other hand, low p-values indicate systematic differences in personal characteristics between the two samples. The merged, unbalanced dataset showed that average household size was significantly different in the two samples. More important, also the share of respondents working in the marine sector differed significantly (p=0.033) and the share of respondents living on the coast differed at a 10% level (p=0.077). It can be argued that living on the coast a person is more likely to care about marine habitat and thus more willing to protect (or the opposite) such habitat (Aanesen et al., 2015). Working in the marine sector means that you more likely will be affected by further protection of CWC and this will affect your WTP for such protection (op cit). Hence, to rule out eventual differences in WTP for protecting CWC due to differences in personal characteristics we used propensity score matching to balance the merged dataset.

The characteristics of the respondents from the male and female datasets, and the corresponding t- and p-values are given in table 2.

Table 2 Characteristics of treatment and control respondents in the balanced dataset, share (%) or mean

	Treatment group (male)	*Control group (female)*	*Test-statistic*	*p-value*
Male share	52.7%	45.1%		0.3736
Age, group 1-6	3.54	3.57		0.8928
Education, level 1-4	3.00	3.044		0.7472
ENGO	12.19%	13.2%		1.00
Urban	74.7%	81.3%		0.3708
Coast	53.85%	54.95%		1.00
Labor force	63.7%	69.2%		0.5299
Marine sector	11%	10%		1.00
Househ. Size	2.12	2.15		0.8414
Income Person, inc.groups 1-10	3.67	4.07		0.215
Income Househ., inc.groups 1-8	3.8	3.9		0.9664
N	91	91		

Table 2 shows that we cannot reject the hypothesis that the respondents are drawn from the same distribution regarding any of the personal characteristics. Hence, by balancing the dataset we have ruled out the possibility that differences in preferences are due to differences in respondent characteristics. The match between respondents in the two samples is poorest when it comes to the share of male respondents, share living in urban areas, and personal income.

3 Results

Estimating the model separately on the two datasets indicates that there may be some differences (see table A2 in the appendix). The respondents in the first dataset (moderated by a middle aged female) are more likely to choose the SQ compared to respondents in the second dataset (moderated by a young male), as the former is above 25% whereas the latter is below 20%. Furthermore, whereas the attributes “attractive for fisheries” and “attractive for oil” are significantly different from zero in the first dataset, this is not the case in the second dataset. These differences, however, may also be due to non-comparable samples and/or systematic preference heterogeneity across the two datasets due to e.g. difference in time and geography when sampling.

To avoid differences in responses due to different sample characteristics, we balanced the merged dataset by the use of the PSM technique (Dehejia and Wahaba, 2002) and ran the models on the merged, balanced dataset consisting of 92 respondents from each of the original datasets.

Running two versions of the MNL model; one with endogenous relative scale and one with interviewer as explanatory variable, yielded both a significant scale parameter and significant interviewer parameter (see tables A3 and A4 in the appendix). These effects, however, may also be caused by differences in other circumstances in the gathering of the two samples, e.g. time and geographic spread, and cannot be attributed solely to the differences in interviewer characteristics. Running these models we used R version 3.2.0. (The R Foundation for statistical Computing).

In order to disentangle the various sources of differences in scale across the two datasets, we ran the MNL model when allowing the relative scale to depend on interviewer, respondent gender and their interaction. The results are shown in the third column of table 3. The models are run using Ox version 7 (Doornik, 2007). They show that interviewer contributes to explain why the scale differs between the two datasets, whereas respondent gender and the interaction between interviewer and respondent gender do not. The results suggests that the relative scale between datasets is 1.413, meaning that the variance in the first group (middle aged female moderator) is roughly 71 % of the variance in the second group (young male moderator).

Still, there may be interactions between interviewer and respondent gender which affect the attribute estimates (stated preferences), but not in a systematic way as assumed by the relative scale parameter. To test for this we applied the MNL model, but let the attribute parameters depend on interviewer, respondent gender, and the interaction between the two. The results are given in the second column of table 3. The formulation of the attribute parameters in the model implies that the (default) reported attribute estimates assume a middle aged female interviewer and a female respondent. The “male” estimate then gives the change in the mean attribute estimate when respondent gender changes from female to male, whereas the “interviewer” estimate gives the change in the mean attribute estimate when the interviewer changes from middle aged female to young male. Finally, the “male*interviewer” estimate gives the change in attribute estimate when respondent gender changes form female to male and interviewer from middle aged female to young male. This means that neither respondent gender nor interviewer has effects on the attribute estimate across all attributes. The attributes small size, large size and habitat are affected when we change moderator from middle aged female to young male. E.g. the attribute habitat, with a mean estimate of 0.952 gets an addition of 0.242 when the interviewer changes to young male. The attributes large size, attractive for oil and attractive for fisheries are affected when the respondent gender is changed from female to male. For example, the mean estimate for the attribute large size becomes 0.216 smaller when the respondent is male instead of female (and the interviewer is a middle aged female). Table 5 shows that changing both respondent gender and interviewer does not have significant effects on the mean attribute estimate for any of the attributes.

Table 3 *MNL models with attributes or scale depending on interviewer and respondent gender. Estimated means (st.errors)*

Attribute	*Attribute estimate*	*Scale estimate*
Small size	-0.0147 (0.083)	-0.462 (0.05)
- male	-0.129 (0.114)	
- interviewer	0.294 (0.206)*	
- male*interviewer	-0.204 (0.284)	
Large size	0.215 (0.084) ***	0.179 (0.052) ***
- male	-0.216 (0.116) **	
- Interviewer	0.478 (0.209) **	
- male*interviewer	-0.167 (0.29)	
Attractive for oil	0.127 (0.053) ***	0.051 (0.031) *
- male	-0.098 (0.073) *	
- interviewer	-0.054 (0.117)	
- male*interviewer	-0.016 (0.169)	
Attractive for fisheries	0.073 (0.055) *	0.113 (0.034) ***
- male	0.156 (0.076) **	
- interviewer	-0.067 (0.144)	
- male*interviewer	-0.137 (0.205)	
Habitat	0.952 (0.061) ***	0.976 (0.055) ***
- male	0.011 (0.084)	
- interviewer	0.242 (0.132) **	
- male*interviewer	0.122 (0.189)	
Cost	-0.627 (0.08) ***	0.627 (0.055) ***
- male	- 0.059 (0.111)	
- interviewer	0.212 (0.175)	
- male*interviewer	-0.697 (0.259)	

Scale		
- male		-0.082 (0.074)
- interviewer		0.3463 (0.089) ***
- male*interviewer		-0.016 (0.128)
Log Likelihood	-5862	-5889
Adj. R2	0.0922	0.0903
AIC	11771	11796
N, K	5902, 24	5902, 9

Alternatively, we may divide the dataset into four subsets, according to interviewer and respondent gender, and then run the model on each of the subsets and compare mean attribute estimates. Irrespective of interviewer, there are differences in stated preferences between female and male respondents. First, female respondents are more likely to choose protection of CWC than are men. Second, female respondents assess the attributes of the protection scenario more evenly than men, e.g. by valuing all attributes (except cost) positively. Men state stronger likes and dislikes, e.g. by valuing some attributes negatively and some positively.

Introducing interviewer characteristics yields a more heterogeneous picture (remains to test for significance of differences). With the young male interviewer the respondent-gender differences become less polarized. Still, female respondents are more likely to choose protection than male respondents, but the difference is smaller compared to with a middle aged female interviewer. With a middle aged female interviewer female respondents assessed both a small and a large increase in protected area positively, whereas men assessed both negatively. With a young male interviewer all respondents assessed a large increase in protected area positively, whereas a small increase was insignificant. Also, with a male interviewer the attributes “attractive for fisheries” and attractive for the oil industry” were insignificant for all respondents, independent of gender. With a female interviewer “attractive for fisheries” was insignificant for male respondents, whereas “attractive for the oil industry” was insignificant for female respondents. The results are given in table A5 in the appendix.

4 Discussion

Previous studies of the effects of interviewer characteristics show that respondents do respond differently to interviewers with different personal characteristics, such as gender, race, dressing code Kane and Macaulay, 1993, Bateman and Mawby, 2004, Loureiro and Lotade, 2005, Gong and Aadland, 2011). Neither of these studies manage to demonstrate that the change in stated preferences due to change in interviewer characteristics depends on personal characteristics of the respondent. Our results also demonstrate that respondents' stated preferences change due to changes in personal characteristics of the interviewer. WE also demonstrate that they change due to changes in respondent characteristics (gender), but like the previous studies, we fail to demonstrate that changes in stated preferences when interviewer characteristics change, depend on respondent characteristics (gender).

Our findings are seemingly at odds with e.g. those of Gong and Aadland (op cit.). They showed that respondents stated higher WTP when interviewed by white and female interviewers compared to non-white and male. The results from our study show that respondents are more willing to protect CWC when interviewed by a young male compared to a middle aged female. Interpreted in a social desirability context this could be explained by more sympathy and a higher propensity to please the younger male interviewer compared to the older female interviewer. With this interpretation our results are in concert with those of Gong and Aadland (2011), if we interpret social desirability to mean sympathy for "non-dominant" characteristics such as younger vs older, female vs male, non-white vs white. In this interpretation, however, it is the age of the interviewer, rather than the gender, which is focal. If we, on the other hand, concentrate on gender, our results are at odds with those of e.g. Gong and Aadland (2011).

A few studies have focused on whether changes in stated preferences due to change of interviewer also depend on respondent characteristics. Neither Kane and Macaulie (1993), Fuchs (2009) nor Gong and Aadland (2011) could demonstrate significant effects when taking into account the respondent gender when analyzing interviewer gender effects. Irrespective of interviewer our data shows that male respondents are less likely to choose CWC protection, and have lower marginal WTP to pay for the different attributes of such protection. This pattern is amplified when interviewer changes from young male to middle aged female. With a middle aged female interviewer the difference between male and female respondents' valuation of changes in the attributes becomes more extreme compared to with a young male interviewer. As an example, when interviewed by a middle aged female, male respondents express a negative WTP for the size attribute, but positive WTP for the habitat and commercial attributes whereas female respondents express positive WTPs for all attributes (except cost). When interviewed by a young male all respondents express positive

WTP for the (large) size attribute and the habitat attribute, whereas the WTP for the commercial attributes are insignificant (but it remains to verify if these results are statistically significant). Hence, the middle aged female interviewer triggers more extreme stated preference differences between women and men than does the young male interviewer. Whether this is due to seniority or gender is not possible to say. One explanation is, as above, that the respondents to a larger degree “restrict” themselves when faced with a younger (male) interviewer, i.e. that the social desirability effect is larger.

In addition to different interviewers, the two datasets are also characterized by being sampled at different times and with different geographic spread, all of which may result in systematic preference heterogeneity. Both the interviewer dummy and the scale parameter are significant, verifying that there are indeed systematic differences in preferences across the datasets. Although tempting to explain this difference by different interviewer, we must admit that it could also be due to different time of sampling or different geographic spread. The male dataset is collected exactly one year later than the female dataset, and there are no obvious reasons for why peoples’ preferences should have changed during this year. There was no focus on CWC in media and economic conditions for people in Norway did not change significantly during this year. The geographic spread in the male dataset is lower than in the female dataset, and covered 4 different municipalities in the south of the Norway. Compared to 22 sampled municipalities covering the whole country, this factor may cause systematic preference heterogeneity across the datasets. But according to tests of the personal characteristics of the respondents, the geographic spread of the respondents in the two datasets does not differ significantly (see table 2). Hence, it can be argued that a major part of the difference in preferences across the datasets is due to different interviewer.

Which implications may these results have for the use of stated preferences surveys and their use in decision making? The unambiguous message is that at least in f2f (and phone) surveys personal characteristics of the interviewer do affect responses. The more ambiguous message is how they affect these responses? The fact that interviewer characteristics affect survey responses in valuation surveys could be exploited by stakeholders interested in specific results from stated preference surveys. Following the results from the present paper this implies that if stakeholders have an interest in highest possible marginal WTPs they should go for a young male interviewer rather than an older female interviewer. Such strategic considerations are of course unfortunate. On the other hand, the fact that results from different experiments and tests on interviewer effects are ambiguous makes such strategic behavior w.r.t valuation surveys more difficult.

The unambiguous message that interviewer characteristics do affect respondents' stated preferences in valuation surveys, be it contingent or discrete choice experiments, may be seen as an argument for web-surveys. In a web-administrated survey there is no person involved who could influence the respondents choices. On the other hand, these surveys come with problems of their own, and there is a growing literature on web-surveys (see e.g. Lindhjem and Navrud, 2011, Baker et al., 2010, Sandorf et al., in prep.).

Conclusions

This paper adds to the rather small literature on interviewer effects in stated preferences surveys. Papers on interviewer effects in valuation surveys to a large degree conclude that such effects are present. Respondents react to personal characteristics of the interviewer such as gender, race, and dressing code. We test for interviewer effects in the form of gender and age by comparing result from a discrete choice experiment implemented as a valuation workshop, moderated first by a middle aged female and next a young male. We find that respondents when interviewed by a young male choose the SQ less frequently and have higher WTP for CWC protection.

Based on previous studies and the present paper the hypothesis that respondents in valuation surveys are sensitive to the appearance of the interviewer seems to have some support. How such interviewer effects work are, however, still ambiguous. Whereas Gong and Aadland (2011) report that white and female interviewers in a phone survey generate higher WTP compared to non-white and male interviewers, our result indicates that a middle aged female interviewer in a valuation workshop generate lower WTP compared to a young male interviewer. The good valued is in the first case a well-known market good (fee for recycling of waste) and in the other case an unfamiliar non-market good (WTP for protection of cold-water coral).

It is not unlikely that the appearance of an interviewer affects the responses of the respondents, and the crucial question is whether this fact renders "personal" stated preferences surveys useless? The introduction and diffusion of web-surveys may be a solution to the interviewer-effect. On the other hand, in some circumstances there is a need for providing the respondents with certain information, to give them the possibility to discuss the good to be valued or ask questions in order for them to be able to give their answers. These are situations where we either want to elicit citizens' preferences (valuation) or where the good to be valued is very complex or unknown such that it is dubious whether information via the net and as part of a web survey will give the respondents a clear understanding of what they are going to value. These are typical situations where the deliberative monetary valuation techniques are recommended, which presuppose the presence of a moderator/interviewer.

References

West, B.T., F.Kreuter, U.Jaenichen (2013) “Interviewer” Effects in face-to-face surveys: A function of sampling, Measurement Error or Nonresponse? Journal of Official Statstics, 29 (2) 277-297

Fuchs, M. (2009) Gender-of-Interviewer effects in a video-enhanced web survey. Social psychology, 40 (1) 37-42

Flores-Macias, F., C.Lawson (2008) Effects of interviewer gender on survey responses: findings from a household survey in Mexico, International journal of public opinion research, 20 (1)

Fosså,

Huddy, L.,m J.Billig, J.Bracciodieta, L.Hoeffler, J.Moynihan, P.Pugliani (1997) The effect of interviewer gender on the survey response, Political Behavior, 19 (3), 197-220

Kane, E.W., L.J.Macaulay (1993) Interviewer gender and gender attitudes, The public opinion quarterly, 57 (1) 1-28

List, J.A., R.P.Berrens, A.K.Bohara, J.Kerkvliet (2004) Examining the role of social isolation on stated preferences, American Economic Review, 94 (3), 741-752

Bateman, I.J., J.Mawby (2004) First impressions count: interviewer appearance and information effects in stated preference studies, Ecological Economics, 49, 47-55

Johnston, R.J. (2006) Is hypothetical bias universal? Validating contingent valuation responses using a binding public referendum, Journal of Environmental Economics and Management, 52, 469-481

Leggett, C.G., N.S.Kleckner, K.J.Boyle, J.W.Duffield, R.C.Mitchell (2003) Social desirability bias in contingent valuation surveys administered through in-person interviews, Land Economics, 79 (4) 561-575

Groves, R.M., N.H.Fultz (1985) Gender effects among te4lephone interviewers in a survey of economic attitudes, Sociological methods and research 14, 31-52

Loomis, J., L.Ellingson, A.Gonzales-Caban, A.Seidl (2006) The role of ethnicity and language in contingent valuation analysis. A fire prevention policy application, American journal of Economics and Sociology, 65 (3), 559-586

Loureiro, M.L., J.Lotade (2005) Interviewer effects on the valuation of goods with ethical and environmental attributes, Environment and Resource Economics, 30, 49-72

Davis, D.W., B.D.Silver (2003) Stereotype threat and race of interviewer effects in a survey on political knowledge, American journal of political science, 47 (1) 33-45

Dehejia, R.H., S.Wahba (2002) Propensity score-matching methods for nonexperimental causal studies, The review of economics and Statistics 84 (1), 151-161

Austin, P.C. (2011) An introduction to propensity score methods for reducing the effects of confounding in observational studies, Multivariate behavioral research 46, 399-424

Spangenberg, J.H. and J.Settele (2010) Precisely incorrect. Monetising the value of ecosystem services, Ecological complexity 7, 327-337

Revelt and Train (1998)

Appendix

Table A1 *Respondent characteristics in the two datasets*

	Middle aged female interviewer		*Young male interviewer*		*p-value*
Male share	53.5%		50%		0.9579
Age	<26y:11%	>65Y:18%	<26y:12%	>65y:20%	0.5028
Education	Primary:7%	Higher:33%	Primary8.5%	Higher:42.5%	0.2785
ENGO	10%		11%		0.8309
Labor market	63%		61%		0.6065
Marine sec	7%		14%		0.0327
Coast	62.7%		58.8%		0.077
Urban	73%		73.5%		0.9691
HSize	2.56		2.09		0.0002
IncPers	<500k	>500k	<500k	>500k	0.166
	77%	23%	70%	30%	
IncHouse	<500k	>1 mill	<500k	>1mill	
	66%	15%	62%	20%	0.8934
N	397		106		

Table A2 *Results from original datasets with middle aged female and young male interviewers*

Attributes	*Middle aged female interviewer*	*Young Male interviewer*
Small size	-0.0835 (0.0566)	0.0803 (0.1283)
Large size	0.1002 (0.0577) **	0.4668 (0.1312) ***
Attractive for oil	0.0748 (0.0365) **	0.0212 (0.0785)
Attractive for fisheries	0.157 (0.0379) ***	0.0143 (0.0946)
Habitat	0.9556 (0.0421) ***	1.2488 (0.0843) ***
Cost	-0.6574 (0.0552) ***	-0.6060 (0.0476) ***
Log likelihood	-4720	-1165
Adj. pseudo R2	0.0737	0.1507
AIC	9451	2342
N, K	4647, 6	1255, 6
Pr(Alt.1)	0.36	0.38
Pr(Alt.2)	0.39	0.41
Pr(SQ)	0.25	0.21

Table A3 *Results from MNL model with interviewer as explanatory variable*

```
Coefficients :
          Estimate Std. Error t-value  Pr(>|t|)
litenX   -0.243832   0.091963 -2.6514  0.008016 **
storX     0.119149   0.090770  1.3127  0.189300
oilnewX   0.046459   0.027792  1.6716  0.094597 .
fishnewX  0.071414   0.031370  2.2766  0.022813 *
hab       1.184668   0.063699 18.5979 < 2.2e-16 ***
cost     -0.545780   0.082602 -6.6074 3.912e-11 ***
interv:1  0.245601   0.112702  2.1792  0.029316 *
interv:2  0.354863   0.111275  3.1891  0.001427 **
```

Table A4 *Results from MNL model with estimated relative scale parameter (the relative scale parameter is parameter 7)*

Balanced dataset

```
Maximum Likelihood estimation
BFGS maximisation, 276 iterations
Return code 0: successful convergence
Log-Likelihood: -1684.013
7  free parameters
Estimates:
      Estimate Std. error  t value   Pr(> t)
[1,] -1.028799   0.090416 -11.3786 < 2.2e-16 ***
[2,] -1.278470   0.108408 -11.7931 < 2.2e-16 ***
[3,] -0.045285   0.020062  -2.2572  0.023992 *
[4,] -0.058210   0.021839  -2.6655  0.007688 **
[5,] -0.685686   0.090121  -7.6085 2.773e-14 ***
[6,]  0.189011   0.058672   3.2215  0.001275 **
[7,] -0.629777   0.074720  -8.4285 < 2.2e-16 ***
```

Table A5 Results from the MNL model for the four subsets

A: Female interviewer - female respondents

```
Frequencies of alternatives:
      3       1       2 
0.19388 0.39796 0.40816 

nr method
4 iterations, 0h:0m:0s 
g'(-H)^-1g = 0.000251 
successive function values within tolerance limits 

Coefficients :
          Estimate Std. Error t-value  Pr(>|t|)    
litenX    0.312793   0.144451  2.1654 0.0303575 *  
storX     0.535183   0.142950  3.7439 0.0001812 ***
oilnewX   0.156690   0.049293  3.1787 0.0014793 ** 
fishnewX  0.019884   0.050762  0.3917 0.6952707    
hab       0.816172   0.113167  7.2121 5.509e-13 ***
cost     -0.338695   0.147379 -2.2981 0.0215547 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -584.5
```

B: Female interviewer - male respondents

```
Frequencies of alternatives:
      3       1       2 
0.32640 0.32017 0.35343 

nr method
4 iterations, 0h:0m:0s 
g'(-H)^-1g = 2.1E-07 
gradient close to zero 

Coefficients :
          Estimate Std. Error t-value  Pr(>|t|)    
litenX   -0.687843   0.167882 -4.0972 4.182e-05 ***
storX    -0.400410   0.162337 -2.4665   0.01364 *  
oilnewX   0.034374   0.062396  0.5509   0.58170    
fishnewX  0.332060   0.066749  4.9747 6.534e-07 ***
hab       1.332417   0.147052  9.0609 < 2.2e-16 ***
cost     -0.541322   0.182084 -2.9729   0.00295 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -470.29
```

C: Male interviewer - female respondents

```
Frequencies of alternatives:
     3      1      2 
0.1487 0.4145 0.4368 

nr method
4 iterations, 0h:0m:0s 
g'(-H)^-1g = 0.000153 
successive function values within tolerance limits 

Coefficients :
          Estimate Std. Error t-value  Pr(>|t|)    
litenX    0.049869   0.181569  0.2747 0.7835798    
storX     0.658137   0.172779  3.8091 0.0001395 ***
```

```
oilnewX   0.021331   0.058317  0.3658 0.7145334
fishnewX  0.024175   0.071763  0.3369 0.7362093
hab       1.388833   0.133260 10.4220 < 2.2e-16 ***
cost     -0.373138   0.171886 -2.1708 0.0299430 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -461.73
```

D: Male interviewer - male respondents

```
Frequencies of alternatives:
      3       1       2
0.24259 0.36111 0.39630

nr method
4 iterations, 0h:0m:0s
g'(-H)^-1g = 1.51E-07
gradient close to zero

Coefficients :
            Estimate  Std. Error t-value  Pr(>|t|)
litenX   -0.06141486  0.16279965 -0.3772   0.70599
storX     0.38163054  0.16027014  2.3812   0.01726 *
oilnewX  -0.00022246  0.06094811 -0.0036   0.99709
fishnewX -0.09108755  0.07440566 -1.2242   0.22088
hab       1.35736796  0.13562338 10.0084 < 2.2e-16 ***
cost     -1.25037425  0.19342194 -6.4645 1.016e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -499.7
```