**Valuing tap water quality improvements using stated preference methods.**

**Does the number of discrete choice options matter?**

**Mikołaj Czajkowski, Marek Giergiczny and Ewa Zawojska**

Faculty of Economic Sciences,
University of Warsaw, Poland

ezawojska@wne.uw.edu.pl

***Selected Paper prepared for presentation at the 2018 Agricultural & Applied Economics Association Annual Meeting, Washington, D.C., August 5-August 7***

# Valuing tap water quality improvements using stated preference methods: Does the number of discrete choice alternatives matter?

Mikołaj Czajkowski*,

Marek Giergiczny

and

Ewa Zawojska

*Faculty of Economic Sciences*

*University of Warsaw, Warsaw, Poland*

* Corresponding author, miq@wne.uw.edu.pl

**Acknowledgments:** Magda Kazimierczuk facilitated this study by providing support in data collection. Mikołaj Czajkowski gratefully acknowledges the support of the Polish Ministry of Science and Higher Education and the Foundation for Polish Science. Ewa Zawojska acknowledges the generous support of the National Science Centre in Poland (Preludium 8, grant no. 2014/15/N/HS4/01328; Etiuda 4, grant no. 2016/20/T/HS4/00013).

# Valuing tap water quality improvements using stated preference methods: Does the number of discrete choice alternatives matter?

**Abstract:** Stated preference methods are widely used for valuing public goods such as tap water quality improvements. Nevertheless, whether respondents reveal their true preferences in surveys is still the subject of academic debate. Existing literature suggests that truthful disclosure of preferences in valuation surveys may pivot on the number of alternatives presented in a single choice task. Although, on a theoretical basis, the use of a single two-alternative choice format (one of which is a "no change" option) has long been recommended because of its potential incentive-compatible properties, empirical studies provide nascent evidence that using more than two choice alternatives may be a more natural setting, because it makes it easier for respondents to find an option that matches their preferences. The lack of consensus about the impact of the number of alternatives on respondents' stated preferences motivates this study. Using data from a discrete choice experiment, we examine whether willingness-to-pay (WTP) estimates depend on the number of alternatives provided in a choice task. We employ a split-sample design that uses two- and three-alternative formats in a stated preference survey of proposed public policies for the improvement of tap water quality (iron and chlorine contents, hardness) in Milanówek, a town in the Warsaw agglomeration in Poland. Drawing on the mixed logit model with scale heterogeneity, we find no significant differences in the mean WTP values elicited with two- and three-alternative tasks, while the WTP estimates based on three-alternative tasks appear to have lower standard errors compared with two-alternative tasks. This finding indicates that the use of three or more alternatives per choice task may offer a way to increase efficiency without significantly biasing the results.

**Keywords:** tap water quality; stated preference methods; contingent valuation; discrete choice experiment; incentive compatibility; number of alternatives

**JEL classification:** Q51, Q25, D12, D82, C25

## 1. Introduction

The quality of tap water supplied to citizens is among the most important public policy issues. Some policies focus on the reduction of health risks associated with the tap water use, while others are mainly concerned with improving the physical characteristics of tap water, which include its smell, taste, color and turbidity. Bearing in mind the high costs related to the implementation of public policies, their correct valuation in terms of benefiting the population is essential.

The most widely used approaches for estimating the value of better tap water quality include the averting behavior method and the contingent valuation technique. The former evaluates the proposed improvement based on the expenditures incurred by consumers who seek to avoid the negative consequences of poor tap water quality, such as the costs associating with buying bottled water or installing water filters (e.g., Um *et al.*, 2002; Dupont and Jahan, 2012). The latter technique infers the improvement value from consumers' choices that are stated in specially designed surveys, which present the respondents with hypothetical policy scenarios and ask them to choose the variant that they prefer most (e.g., Adamowicz *et al.*, 2011; Polyzou *et al.*, 2011). Here we use the term "contingent valuation" in line with the nomenclature in stated preference research, as suggested by Carson and Louviere (2011); "contingent valuation" is independent of the preference elicitation method, so it encompasses both open-ended questions and close-ended choice questions (including discrete choice experiments).

Although widely applied, each of these techniques has some limitations, which calls into question the potential of their practical application for policy purposes. An important shortcoming of the averting behavior method is that consumers' actions that are interpreted as defensive may be actually motived by other reasons, for example, a personal preference for bottled water because of its convenience. Therefore, the technique is likely to overestimate true values (Bresnahan *et al.*, 1997).

The reliability of the value estimates based on contingent valuation data has been questioned because of the method's potential lack of incentive compatibility, that is, the contingent valuation surveys are thought not to provide the respondents with the incentives under which their optimal response strategy is to reveal their preferences truthfully. Recent research (Carson and Groves, 2007) provides guidelines for making the contingent valuation surveys incentive compatible, which impose strict restrictions on the survey design. The guidelines suggest that such surveys need to be perceived as consequential by the respondents, meaning that they should believe in the real consequences of the survey's outcomes and care about the policy that is ultimately introduced. Moreover, the surveys need to employ a coercive payment mechanism that distributes a payment to all agents if the policy is implemented. Finally, the guidelines require that the surveys employ a two-alternative task format with one of the two choice options being a "no change" alternative (status quo). Because all the theoretically suggested requirements are often not fulfilled in actual contingent valuation studies, the validity of the method itself has been questioned (Zawojska and Czajkowski, forthcoming).

As follows from the Gibbard-Satterthwaite theorem (Gibbard, 1973; Satterthwaite, 1975), no question format providing more than two choice alternatives is incentive compatible unless extra constraints are imposed on the respondents' preferences (Carson and Groves, 2007). Although the incentive-compatible properties of a two-alternative choice task are well known, researchers often employ extended formats with multiple alternatives per choice task because of statistical efficiency gains – more information is revealed by respondent's single choice. Thus, for the practical application of the contingent valuation method, the fundamental question is whether the use of the multiple-alternative task format results in biased value estimates and, if so, how large this bias is. Knowledge of the magnitude of the bias could help applied researchers make trade-offs between the incentive compatibility and statistical efficiency of the survey instrument. Note that the incentive compatibility conditions also rule out using more than one choice task per respondent. This issue has also been a subject of extensive research (e.g., Carlsson *et al.*, 2012; Day *et al.*, 2012; Czajkowski *et al.*, 2014) and is not pursued here.

Our study employs the contingent valuation method with the elicitation format of a discrete choice experiment (Carson and Czajkowski, 2014) to measure the economic benefits of having high-quality tap water. We empirically examine whether the number of alternatives provided in a choice task matters for the preferences stated in contingent valuation surveys. We employ a split-sample design that uses two- and three-alternative task formats in a field survey involving the valuation of proposed public policies for tap water quality improvement in Milanówek, a town in the Warsaw agglomeration in Poland.

Drawing on the mixed logit (MXL) model with scale heterogeneity, we find no statistically significant differences in the mean value estimates based on two- and three-alternative choice data. The value estimates derived from the three-alternative choices appear to have smaller standard errors than do the estimates derived from the two-alternative choices. As a result, we argue that using more than two alternatives per single task may offer a way of increasing statistical efficiency without significantly biasing the results.

The lack of significant differences in the value estimates derived from two- and three-alternative choices indicates that respondents do not necessarily partake in strategic misrepresentation of their preferences in multiple-alternative task formats. We point to the two possible causes underlying this result – task complexity and uncertainty with respect to other respondents' preferences. The respondents might find it difficult to determine which policy alternative is preferred in the population for each comparison of a few alternatives, or they might not have enough information about how the preferences are distributed in the population and how the choices made in the survey will translate into a final decision.

The paper is structured as follows. Section 2 reviews the existing literature with respect to (1) the studies devoted to the valuation of tap water quality and (2) the empirical evidence regarding the effect of the number of alternatives on truthful preference disclosure. Section 3 discusses the research methodology by providing details on the survey design for contingent valuation and its implementation and by

describing the econometric approach used to model consumers' preferences. Section 4 reports the results of our field study and analyzes the practical importance of the incentive-compatible condition requiring two-alternative choice task formats. Section 5 concludes and summarizes our findings.

## 2. Literature review

### 2.1. Valuation of tap water quality improvements

Tap water quality improvements typically concern the reduction of health risks associated with its use or the amelioration of its physical characteristics, such as smell, taste and turbidity. Starting with the former, tap water quality changes aimed at mitigating the probability of health problems related to its use are usually valuated because of observed deterioration in water quality or the occurrence of certain accidents that suggest water contamination. For example, finding trichloroethylene, a volatile synthetic organic chemical, in one of the borough's wells in southeastern Pennsylvania prompted Abdalla *et al.* (1992) to estimate a reduction in economic benefits due to the increased defensive expenditures, including bottled water purchases and the installation of home filtration systems. The detection of bacteria, minerals and organic material in drinking water supplies underlay the averting cost analysis by Collins and Steinback (1993). Phenol spills in Seoul, Korea, which had badly polluted the water there, and Koreans' subsequent resistance to use tap water motived the valuation studies of Kwak and Russell (1994) and Um *et al.* (2002).

Studies employ various techniques to estimate the value of better tap water quality. In the context of the valuation of health risk reduction, the averting behavior method is particular important because it enables the determination of the value of goods and services based on consumers' averting (or defensive) actions. The idea underlying the method is that the valuation of something "bad" can be measured by the expenses incurred to avoid this "badness". In the face of increased health risks resulting from their exposure to poor-quality tap water, consumers undertake various actions to defend themselves against this risk, such as boiling water, purchasing bottled water or installing home water treatment systems, which constitutes the basis for estimating averting expenditures (Harrington *et al.*, 1989; Abdalla *et al.*, 1992; Laughland *et al.*, 1993; Whitehead *et al.*, 1998; Dupont and Jahan, 2012).

One of main criticisms of the averting behavior method is that the motivations behind consumers' actions that are interpreted as defensive may arise from other reasons. For example, consumers may simply prefer bottled water because of its taste or convenience. Consequently, the value of tap water improvements estimated based on the reported expenditures is likely to be exaggerated (Bresnahan *et al.*, 1997). Abrahams *et al.* (2000) address this issue empirically and show the potential problem of overstating estimates based on the averting behavior method. Drawing on data for Georgia in the U.S., consumers appear to purchase bottled water for health- and non-health-related reasons. Thus, the

researchers suggest an approach that adjusts defensive expenditures to exclude non-health-related components.

The contingent valuation method offers an alternative approach to valuate tap water quality improvements. This technique is a so-called stated preference method because it relies on consumers' choices, as stated in specially designed surveys, which typically describe hypothetical scenarios for improvement policy and later ask consumers to express their preferences regarding whether the proposed policy should be introduced. The data on consumers' choices in a survey allows for the formal modeling of their preferences and, thereby, an estimation of the value of the proposed policies to consumers.

In the context of reducing the health risks associated with the use of tap water, Kwak and Russell (1994) use the contingent valuation technique to estimate the value of lowering the probability of phenol accidents, which happened in Korea due to tap water contamination, for Seoul residents. The method was also applied by Cho *et al.* (2005), who valuate iron and sulfate reductions in community water supplies in southwestern Minnesota, and by Adamowicz *et al.* (2011), who elicit consumers' WTP to mitigate risks of microbial and cancer illnesses and deaths related to the use of tap water.

The value of tap water quality improvements, which enhance its physical characteristics, is usually assessed with the contingent valuation technique, which enables the valuation of hypothetical states, which have not (yet) taken place. Some research focuses on improvement in general; for example, Polyzou *et al.* (2011) elicit consumers' WTP for a better city system for pipe water that would improve tap water quality. However, other studies examine the value of separate attributes of the proposed policy improvements. For instance, Scarpa *et al.* (2012) and assess the value of changes in chlorine odor, chlorine taste, water turbidity and calcium carbonate stains, while Day *et al.* (2012) examine how much improvements with respect to water color, chlorine smell and taste are valued by consumers.

Yoshida (2009) combines the valuation of tap water improvements related to taste and safety, including the reduction of chlorine and trihalomethane and the replacement of old pipes, in a single study. The research employs the averting behavior and contingent valuation techniques and shows that the former produces significantly higher value estimates than the latter. In the context of the reduction of health risks related to tap water use, Gnedenko *et al.* (2000) note a similar relationship between the value estimates derived from the two techniques. These findings substantiate the previously mentioned drawback of the averting behavior method, i.e., that not all actions that are interpreted as defensive are indeed motivated by defensive concerns.

The existing literature also provides research that assesses better tap water quality in general, without focusing on particular characteristics of the proposed improvement. Employing the contingent valuation method, Genius *et al.* (2008) examine the value of ameliorating tap water quality in Rethymno on the

Greek island of Crete and find that the local residents are willing to pay 17.64% more than the average water bill to implement the change. A contingent valuation study of Awad and Holländer (2010) reveals that Palestinian households would annually pay 627 NIS more for tap water quality improvement; however, the authors emphasize that the estimated amount seems unlikely to be realistic, given the obtained coefficients and the marginal impacts of key socio-demographic factors that diverge from the economic theory and the existing evidence. Kwak *et al.* (2013) assess the value of tap water quality improvement for residents of Pusan, Korea, which is found to constitute, on average, 36.6% of their monthly water bill. Drawing on data for 10 OECD countries, Beaumais *et al.* (2014) observe that, for the pooled sample, households are, on average, willing to pay 7.5% more than the median annual water bill to improve the tap water quality.

The empirical evidence suggests that consumers' socio-demographic characteristics, as well as their risk perceptions and environmental concerns, impinge on valuation of better tap water quality. Studies consistently report a positive effect of income (Kwak and Russell, 1994; Traore *et al.*, 1999; Gnedenko *et al.*, 2000; Genius *et al.*, 2008; Polyzou *et al.*, 2011; Beaumais *et al.*, 2014) and the presence of small children in the family (Abdalla *et al.*, 1992; Genius *et al.*, 2008) (Luzar and Cosse, 1998) on the estimated values. On the other hand, older respondents appear to valuate tap water quality improvements to a lesser degree than younger respondents do (Abrahams *et al.*, 2000; Gnedenko *et al.*, 2000; Um *et al.*, 2002; Beaumais *et al.*, 2014). With respect to the impact of education and gender, the existing literature provides mixed evidence. While Beaumais *et al.* (2014), Cho *et al.* (2005) and Whitehead *et al.* (1998) find that more educated individuals are willing to pay more for better tap water quality, the study of Janmaat (2007) reveals an opposite relationship. Similarly, there is no consensus regarding whether females are inclined to pay more (Gnedenko *et al.*, 2000; Genius *et al.*, 2008) or less (Beaumais *et al.*, 2014) than males. Scarpa *et al.* (2012) study gender-based differences in taste with regard to better tap water quality. Although the null hypothesis of no difference is rejected, in terms of economic magnitude, the differences appear small and individually significant only for some attributes of the tap water improvement policy.

Personal attitudes also play a role in the valuation of tap water quality improvements. Individuals who are more concerned about environmental degradation are inclined to pay higher amounts for better tap water quality (Traore *et al.*, 1999; Beaumais *et al.*, 2014). The estimated willingness-to-pay values also appear to be higher for respondents who perceive that current tap water quality is poor (Whitehead *et al.*, 1998; Gnedenko *et al.*, 2000; Um *et al.*, 2002; Janmaat, 2007). In particular, the importance of individual perceptions is raised by Um *et al.* (2002). Drawing on an averting behavior study, the authors find that the value estimates for reducing tap water pollution from the current level to an objectively acceptable level are considerably lower than the value estimates for reducing tap water pollution from the current level to a subjectively perceived drinkable level.

### 2.2. Empirical evidence on the role of the number of alternatives on stated preferences

The existing literature provides mixed evidence about the impact of the number of choice alternatives presented in a choice question on the preferences elicited in contingent valuation surveys.

From a theoretical perspective, the incentive-compatible properties of a single two-alternative question format have long been known; therefore, its use has been recommended in stated preference valuation studies (Arrow *et al.*, 1993). It constitutes a fundamental finding in mechanism design theory: any response format providing a choice between more than two alternatives cannot be incentive compatible unless additional constraints are imposed on the respondents' preferences (Gibbard, 1973; Satterthwaite, 1975). Hence, drawing on the theory, truthful preference disclosure can be guaranteed only if a valuation question in a stated preference study is single and employs a two-alternative response space. Vossler *et al.* (2012) further develop the theory and show that under certain conditions a sequence of two-alternative questions is also incentive compatible.

There is some empirical evidence that corroborates the conclusions suggested by mechanism design theory. In the context of private goods, Xu *et al.* (2013) experimentally observe that participants who have three choice alternatives per question do not choose the variant that is unconditionally their most preferred. If they believe that their preferred alternative has the least chance of winning in the plurality voting, agents predominantly choose their second choice. Rose *et al.* (2009) also report bias in respondents' stated preferences that is attributed to the number of alternatives. Interestingly, they find two contradicting directions of the bias – for Australian and Taiwanese respondents, the value of travel-time savings is increasingly overstated as the number of alternatives per single question rises, whereas for Chilean respondents, the declared value of travel-time savings decreases when the number of alternatives increases. Based on this evidence, Rose *et al.* (2009) claim that the impact of this survey design dimension on value estimates may be local and not directly transferable across countries and cultures. Further, Liu *et al.* (2014) conduct a convergent validity test of a multiple-alternative choice question versus a two-alternative choice question and find that the welfare estimates obtained from data from the two question formats differ significantly, and, moreover, the value distributions derived from the multiple-alternative data are more dispersed than those from the two-alternative data.

Although mechanism design theory postulates the use of the two-alternative format because of its incentive compatibility, some empirical studies show a better match between stated and actual preferences when more than two alternatives are provided for a single choice question. In an induced-value experiment, Collins and Vossler (2009) observe more deviations in respondents' choices from their optimal option in two-alternative questions than in three-alternative questions. Similarly, drawing on research in an out-of-laboratory context, Rolfe and Bennett (2009) conclude that more robust models can be estimated on data from three-alternative choice data compared with two-alternative data. Each choice question in their research included a "not sure" option in addition to the two or three alternatives.

The authors observe a considerably higher rate of "not sure" responses when respondents choose between two alternatives, which they attribute to respondents' greater ease in choosing between three than two alternatives. When a broader range of alternatives is provided, survey participants are more likely to find the option that matches their preferences best.

Answers on a mailed survey reported by Ready *et al.* (1995) reveal lower (although not statistically significantly lower) value estimates derived from two-alternative choice data compared to a single multiple-choice question with six alternatives characterized by various certainty levels, ranging from "definitely yes" to "definitely no" responses. The result is not surprising because it is more difficult to say exact "yes" in a two-alternative question, whereas the other format allows for less unequivocal answers, such as "probably yes" or "maybe no". Thus, the broader response space may capture respondents' true preferences more accurately, the reason being that they can find options that better reflect their preferences regarding the good being valued.

Hensher (2006a) argues in favor of the use of more complex choice questions, including more than two alternatives, because "relevance is more important than trying to limit cognitive burden." Real-life choices are complex and often involve far more alternatives than are provided in a single choice question. Therefore, if an option that a respondent would most likely choose in a certain context is not available, then they might be more likely make an error or choose the status quo alternative, which leads to a potential loss of information from the perspective of the researcher.

On the other hand, increasing the number of alternatives – and hence the question's complexity – may prompt respondents to use simplifying heuristics and make them defer or avoid choices (Dhar and Simonson, 2003). Adamowicz *et al.* (2006) find that respondents are more likely to choose a status quo option in a three-alternative question than in a two-alternative question, however, they do not explain the result, nor do they investigate it further.

The two contradicting effects of the number of alternatives on consumers' behaviors are discussed by Zhang and Adamowicz (2011) in their empirical study that compares preferences revealed in two-an three-alternative choice tasks. The authors conclude that task complexity increases the probability of choosing the status quo option, while enhanced preference matching in three-alternative choice tasks decreases this probability.

Further, the two effects are elaborated in a theoretical model by Kuksov and Villas-Boas (2010). Although in the context of an actual consumer decision, not in the context of contingent valuation, the authors analyze the impact of an increase in the number of alternatives with respect to search and evaluation costs. According to their line of reasoning, numerous choice options, which are related to high decision complexity, force consumers to involve themselves in intensive search to find a satisfactory variant among the available alternatives. This search may appear too costly and, in turn,

may prompt them to avoid making a choice. By contrast, when the number of offered alternatives is (too) low, consumers might not have enough hope that they will find satisfactory fit, and they, in turn, might not be inclined to engage in the search at all and avoid making a decision. These two opposing forces lead the authors to conclude that a finite optimal number of alternatives exists. The crucial question concerns how much the number of alternatives can be increased to reinforce the accuracy of respondents' stated preferences.

The problem of the optimal number of alternatives for minimizing the estimated utility function error term variance (or scale) has been addressed empirically. DeShazo and Fermo (2002) examine the issue in an on-site survey regarding services and infrastructure in new national parks, using question formats that differ in the number of alternatives (in one survey the number of alternatives ranged from two to seven, in another it ranged from six to nine). Their study shows that the variance of the error term in the consumer's utility function follows a U-shaped pattern – up to a threshold number of alternatives, the variance decreases, and, at some point, it starts to increase. The initial decline in the variance is attributed to a better match between true preferences and the provided alternatives, whereas its additional rise results from a considerable increase in question-based complexity. Similar findings are reported by Meyerhoff *et al.* (2014) and Caussade *et al.* (2005), who, based on choice data for valuation questions consisting of three, four and five alternatives, also derive a U-shaped relationship between the number of alternatives and the error-term variance, with the optimal number of alternatives (in the sense of the lowest error-term variance) being four. In addition, Caussade *et al.* (2005) note that, if ranked according to the relative impact on the error-term variance, the number of alternatives belongs to one of the most important dimensions of the survey design.

However, the empirical evidence supporting the existence of an optimal number of alternatives is ambiguous. A field, private good study by Conlon *et al.* (2001) reveals that, for choices with up to seven alternatives, the exactness of the reflection of true preferences in declared decisions is not affected by question-based complexity expressed by the number of alternatives per choice task and the number of attributes per alternative. (The authors report that the difficulty of the trade-offs that respondents need to make is rather caused by varying covariance between attributes.) Some other researchers also do not find a significant role of the number of alternatives in preference disclosure. Eliciting consumers' preferences for various modes of transport for commuting trips, Arentze *et al.* (2003) observe no significant difference in the error-term variance when they compare responses across two- and three-alternative questions, which they attribute to a choice dominance effect – if one mode is strictly preferred over the other two modes, then choosing between the alternatives among which this dominant mode appears is easy compared to cases in which this best mode is not included. Carson *et al.* (2011) also report no significant differences in participants' behaviors between two- and three-alternative questions. Drawing on a laboratory-based, induced-value experiment in a public good context, they conclude that

subjects rarely take advantage of the opportunity to strategically misrepresent their preferences, although the multiple-choice format encourages rational agents to vote for their second best option.

Finally, it is worth noting that research devoted to the role of design dimensionality in stated preference valuation studies shows its impact on declared choices, especially through its influence on variations in stated willingness-to-pay (WTP) values. Hensher (2004) reports, on average, higher values for travel-time savings resulting from the use of more complex designs, including those with more alternatives. However, the impact of design complexity on stated choices is observed to be significantly mitigated, or even eliminated, when the WTP estimates are conditioned on all design dimensions (Hensher, 2006b; Chintakayala *et al.*, 2009). On the other hand, when WTP estimation fails to account for all design dimensions, the WTP values are significantly different for data compared to questionnaires employing various numbers of alternatives per choice set (Hensher, 2006b).

In summary, the problem of selecting the optimal number of alternatives is multi-dimensional, with incentive compatibility (and hence possible bias resulting from using more than two alternatives) being probably the most profound. More research and empirical results obtained in different settings are clearly needed to develop univocal recommendations regarding the correct number of alternatives to use in discrete choice experiment contingent valuation surveys.

## 3. Data and methods

### 3.1. Survey design and implementation

Our survey aims to elicit the preferences of the residents of Milanówek, a town in the Warsaw agglomeration in Poland, regarding proposed public policies for tap water quality improvement. Although tap water in this region fulfils all necessary sanitary norms and is safe for consumption, people often avoid drinking water directly from the tap. Some residents are skeptical about whether the tap water contains harmful substances, while others are unsatisfied with the water's physical characteristics, such as its smell, taste and color. The public policies in our survey propose improving tap water quality by ameliorating its characteristics.

The proposed policy attributes and their levels are reported in Table 1. The respondents were asked to consider the reduction of iron and chlorine concentrations and the reduction of water hardness. The choice of the attributes and their levels was premised on the information collected in individual interviews with the citizens of Milanówek. These attributes cover the most commonly mentioned problems associated with the tap water quality, as indicated by our preliminary qualitative investigation, which were feasible to be improved, according to the municipal office of water provision and sewers.

Groundwater in this area is characterized by high iron concentrations (on average 160 µg/l), which erodes the taste, smell and clarity of tap water and may cause reddish-brown staining. Proposed public policies consider reducing iron by 50%, 75% and 95% with reference to the current state.

Water hardness is another negative attribute of tap water in Milanówek. Its average level is equal to 30°f, which means that the tap water is hard. Its negative consequences are mainly sediment build-up on domestic appliances using tap water and the necessity of using more detergents.

The concentration of chlorine, which is used as disinfectant in tap water in Milanówek, is not high and does not exceed 4 µg/l. This level of chlorine is not noticeable by most consumers; however, individuals who are more sensitive may be able to detect it – for them, the chlorine content might considerably impinge on the water quality, negatively affecting its taste and smell. The considered improvement proposes replacing the current method of disinfection based on the chlorine use with ozonizing, and much lower chlorine levels would only be used to prevent secondary contamination in pipes.

The last attribute of the improvement policy was the cost related to its implementation, which every household would have to bear. The payment vehicle was coercive, in the sense that the cost would be imposed on every household in the form of increased monthly payments, if the improvement policy was implemented.

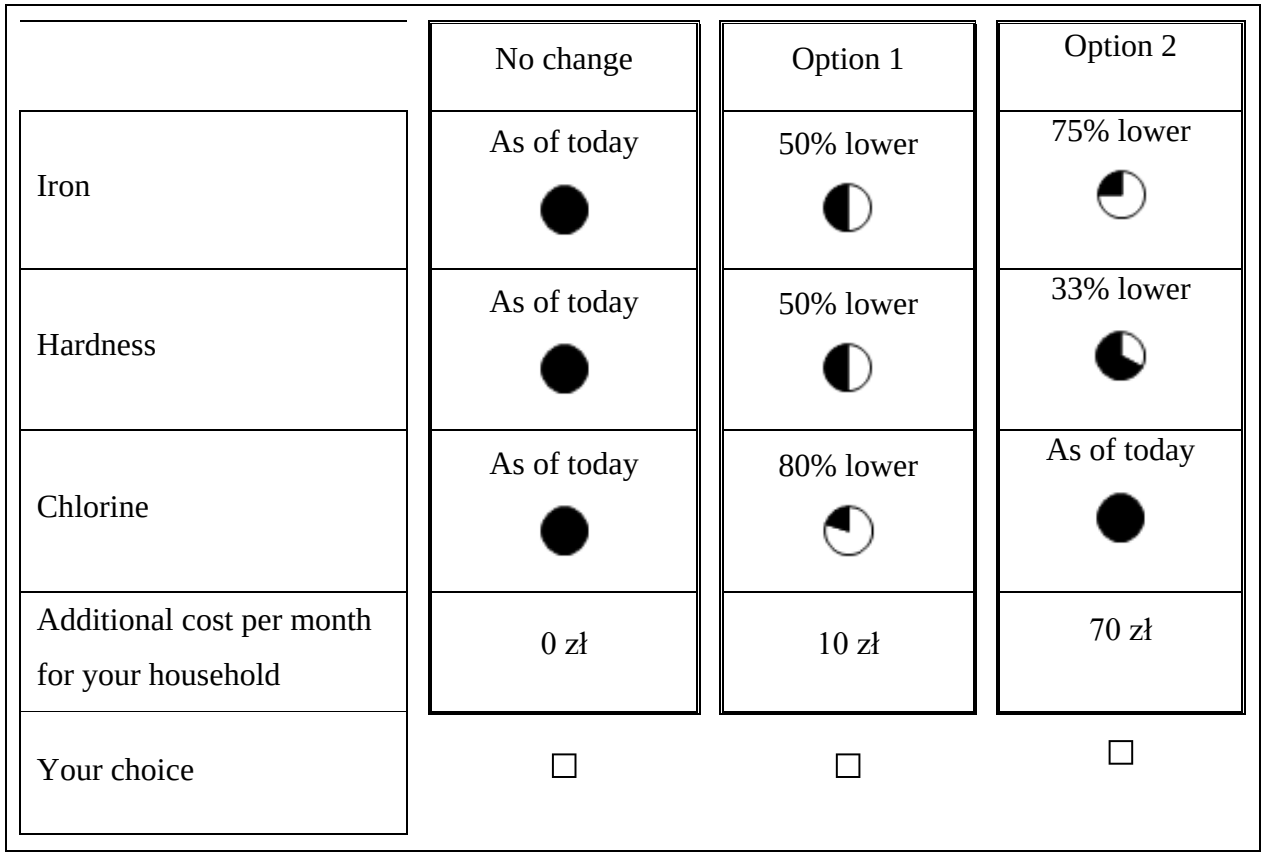
To examine the effect of the number of alternatives on the respondents' stated preferences, we introduced two treatments, which varied only with respect to the number of alternatives provided for a single question. In addition to the status quo option, either one or two policy proposals were included in the choice task. This design allowed us to compare a statistically more efficient design consisting of three alternatives with the theoretically suggested, incentive-compatible two-alternative question. The respondents were randomly assigned to either the two- and three-alternative treatments.

Table 1. Proposed policy attributes

| **Attribute** | **Current level** | **Proposed policy levels** |
|---|---|---|
| Iron content | 160 µg/l | 80 µg/l (-50%), 40 µg/l (-75%), 8 µg/l (-95%) |
| Hardness | 30°f | 20°f (-33%), 15°f (-50%) |
| Chlorine content | 4 µg/l | 0.8 µg/l (-80%) |
| Additional monthly cost for a household | 0 EUR | 2.50, 5.00, 7.50, 10.00, 12.50, 15.00, 17.50 EUR |

Figure 1 presents an example of a choice task from a three-alternative treatment, in which a respondent is asked to make a choice between two possible policies for tap water quality improvement and the status quo. The survey included 12 choice tasks per respondent; depending on the treatment, all choice tasks for the same respondent had either two or three alternatives (namely, in addition to the status quo option, all the choice tasks include either one or two policy options).

Figure 1. An example of a choice task (originally in Polish)




The combinations of attribute levels in every choice task were generated by maximizing the expected Bayesian D-efficiency for a multinomial logit model (Bliemer *et al.*, 2008; Scarpa and Rose, 2008) with priors for the choice parameters obtained from a pilot study and personal interviews. The status quo defined by no change in the attribute levels as compared to today was the same for each respondent; specifically, it was not affected, for example, by a variation in age of the infrastructure or in the groundwater characteristics in different parts of the region because the area of Milanówek is small (13 km2; Central Statistical Office, 2015), and all the dwellings use the same infrastructure and the same water source.

The survey began with "warm-up" questions gauging the respondents' subjective judgements of the tap water quality in Milanówek. This was followed by a detailed description of all of the policy attributes and a typical reminder about budget constraints. A sequence of 12 choice tasks was then presented, where respondents were instructed to choose the alternative that they preferred most from the provided set. At the end of the survey, socio-demographic data were collected.

The survey was designed based on extensive pretesting, including individual interviews with potential survey respondents (verbal protocols), and a pilot study. The survey was administered by mail to households in Milanówek. A screening question was used to restrict the sample to Milanówek residents

only. The data were collected in 2013. Out of the 4,200 surveys delivered to households, 804 surveys were collected (response rate was 19.42%).

### 3.2. Econometric framework

The random utility framework (McFadden, 1974) constitutes the foundation of consumers' preference modelling based on discrete choices. This model assumes that a consumer derives utility from observable characteristics of the proposed policy and unobservable factors that are expressed by a stochastic component. Thus, the utility that consumer *i* obtains from choosing alternative *j* can be formally represented as follows:

$$U_{ij} = \mathbf{a}_i' \mathbf{x}_{ij} + \beta_i c_{ij} + e_{ij}, \quad (1)$$

where $\mathbf{x}_{ij}$ is a vector of non-monetary choice attributes; $c_{ij}$ denotes the monetary attribute; and $e_{ij}$ represents unobservable (from the modeler's perspective) factors that affect consumer's choices. It is important to note that parameters $\mathbf{a}_i$ and $\beta_i$ are consumer-specific, which introduces preference heterogeneity among respondents and leads to a mixed logit model. Instead of these parameters being separately estimated for each respondent, it is typically assumed that they follow a predefined, multivariate distribution. If the parameters are the same for all respondents, preferences are assumed homogenous across the entire sample, resulting in a multinomial logit model (MNL).

The stochastic component of the utility function, $e_{ij}$, has unknown, potentially heteroskedastic variance ($\mathrm{var}\left(e_{ij}\right) = s_i^2$). The model is usually identified upon normalizing the variance, such that the error term, $\varepsilon_{ij} = e_{ij}\, \pi / \sqrt{6} s_i$, is an independent, identically extreme value type one distributed with constant variance ($\mathrm{var}\left(\varepsilon_{ij}\right) = \pi^2/6$). This results in the following specification of the model, called the mixed logit (MXL) model:

$$U_{ij} = \sigma_i \mathbf{a}_i' \mathbf{x}_{ij} + \sigma_i \beta_i c_{ij} + \varepsilon_{ij}, \quad (2)$$

where $\sigma_i$ is a scale coefficient equal to $\pi / \left(\sqrt{6} s_i\right)$. Note that the model specification in (2) represents the same preferences as those given by (1). The estimates $\sigma_i \mathbf{a}_i$ and $\sigma_i \beta_i$ do not have direct interpretations; however, they can be interpreted in relation to each other because the scale coefficient is then canceled out.

Finally, given the interest in estimating consumers' willingness-to-pay values for the non-monetary attributes, it is convenient to introduce the following modification, which is equivalent to a money-metric utility function:

$$U_{ij} = \sigma_i \beta_i \left( \frac{\mathbf{a}_i'}{\beta_i} \mathbf{x}_{ij} + c_{ij} \right) + \varepsilon_{ij} = \sigma_i \beta_i \left( \boldsymbol{\alpha}_i x_{ij} + c_{ij} \right) + \varepsilon_{ij} . \quad (3)$$

The vector of parameters according to the non-monetary attributes in (3), $\boldsymbol{\alpha}_i$, is scale-free and can be directly interpreted as a vector of the implicit values of these attributes. Thus, a model estimated in the form of (3) is called a model in willingness-to-pay space.

To verify whether the number of alternatives per single choice task impinges on stated preferences, we estimate the MXL model in which each attribute interacts with treatment dummies for the two- and three-alternative surveys. This specification allows us to obtain separate willingness-to-pay estimates for two- and three-alternative treatments, thereby testing whether the value estimates differ significantly.

In our study, we not only control the effect of the number of alternatives per choice task on willingness-to-pay estimates but also account for its possible impact on scale (Czajkowski *et al.*, 2015). We incorporate this effect by making the scale dependent on a treatment-related covariate (a dummy for the three-alternative treatment), $z_i$. The model is operationalized by estimating the parameters of the underlying normal distribution, i.e., the scale is estimated as follows:

$$\sigma_i = \exp\left( \bar{\sigma} + \theta' z_i \right) . \quad (4)$$

We are then able to calculate the scale parameter difference between each of the treatments.

The model was estimated using a custom code developed in Matlab, available from https://github.com/czaj/DCE under CC BY 4.0 license. The translation of the original questionnaire and the dataset are made available online at czaj.org.

## 4. Results

Out of the 4,200 mailed surveys, we obtained 403 responses to the two-alternative questionnaire and 401 responses to the three-alternative questionnaire. Table 2 presents the characteristics of the two samples with the statistical tests of significance for the differences between the samples. The two- and three-alternative samples do not differ significantly with respect to any of the socio-demographic characteristics considered.

Table 2. Socio-demographic characteristics of the treatment samples and their equality test results

| | **Two-alternative treatment** | **Three-alternative treatment** | **Test statistics** | **P-value** |
|---|---|---|---|---|
| | **Sample means**[1)] | | | |
| *Years lived in Milanówek* | 32.69 | 32.68 | 77,723 | 0.73 |
| *Age in years* | 51.59 | 51.36 | 73,407 | 0.93 |
| *Household size* | 2.84 | 2.82 | 78,793 | 0.90 |
| *Number of minor children in household* | 0.45 | 0.49 | 77,835 | 0.93 |
| | **Shares** (%)[2)] | | | |
| *Males* | 41 | 36 | 2.22 | 0.14 |
| *Education* | | | 5.15 | 0.16 |
| Elementary | 2.8 | 4.3 | | |
| High school | 33.9 | 32.5 | | |
| Vocational | 18.6 | 13.8 | | |
| Higher | 44.6 | 49.4 | | |
| *Monthly household income* | | | 10.11 | 0.12 |
| Below 2,000 PLN | 14.5 | 22.8 | | |
| 2,000 – 2,999 PLN | 19.6 | 17.3 | | |
| 3,000 – 3,999 PLN | 22.7 | 17.3 | | |
| 4,000 – 4,999 PLN | 13.4 | 12.8 | | |
| 5,000 – 7,499 PLN | 14.8 | 13.9 | | |
| 7,500 – 9,999 PLN | 6.2 | 6.1 | | |
| More than 10,000 PLN | 8.8 | 9.7 | | |

*Notes*: [1)] The equality of the distributions of the characteristics measured on a continuous scale is verified with the Wilcoxon-Mann-Whitney test. [2)] The equality of the proportions of the characteristics measured on a discrete response scale is verified with the Chi-square test.

Table 3 reports the results of three specifications of the MXL model with scale heterogeneity. To ensure the stability of the results, the log-likelihood functions and the standard errors of the estimates are simulated using 10,000 Sobol draws with a random linear scramble and random digital shift (Czajkowski and Budziński, 2015). The estimation results obtained from other, more restrictive models (available from the authors upon request), which include the multinomial logit and the mixed logit model without correlations, and the generalized mixed logit model do not differ substantially from the results presented here, hence our conclusions are not sensitive to the modelling approach employed. We discuss here the results of the MXL model with correlated parameters as it is much more flexible than the other traditional approaches.

To verify the robustness of our findings, and to exclude the possibility that our results are affected by the sequencing issues, the models could be re-estimated using only the data for the first choice task. However, this procedure finds no application in the context of our research which was a mail survey and respondents could move back and forth through the questionnaire. Thus, we conduct the analysis on the full data set which includes respondents' answers to all choice tasks.

Each specification in Table 3 includes an alternative-specific constant for the current situation (status quo) and dummy-coded levels of reductions of iron, chlorine and water hardness (see Table 1 for details). The models are estimated in WTP space, as outlined in the preceding section; thus, every parameter can conveniently be interpreted as marginal WTP for a given attribute (in EUR). The preference parameters are assumed to be normally distributed, while the preference-space equivalent of the parameter of the monetary attribute is assumed to be log-normally distributed.

Model 1 assumes that all preference parameters are equal across the two treatments and that any differences are driven by the scale parameter. Model 2 relaxes the assumption of the same preference parameters across treatments by including interactions of the means of the preference parameters with a dummy variable for the three-alternative treatment. Model 3 further relaxes the assumption of the equality of the preference parameters for the two treatments and assumes that the randomly distributed preference parameters (means, standard deviations and correlations) are treatment-specific.

In every specification in Table 3, all preference parameters are highly significant and have the expected sign, indicating that respondents are willing to pay for the improvement of tap water quality. The magnitudes of the parameters show that chlorine reduction is valued considerably less than the other proposed changes. Taking into account the relatively low baseline level of chlorine concentration and the relatively high iron content in the tap water in Milanówek, this result coincides with the expectations. The coefficients for the reduction of iron concentration and water hardness do not diverge much, which suggests that respondents are, on average, similarly concerned about these two water quality characteristics. The statistically significant coefficients of the standard deviations indicate the presence of preference heterogeneity, which justifies the use of the MXL model specifications.

Table 3. Estimation results of the MXL models in WTP space (denominated in EUR) investigating preference and scale differences between two- and three-alternative treatments

| | **Model 1**<br>**MXL with preference parameters equal for both treatments** | | **Model 2**<br>**MXL with preference parameter means interacting with a treatment dummy** | | | **Model 3**<br>**MXL with treatment-specific preference parameters** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Two-alternative treatment | | Three-alternative treatment | |
| *Preference parameters* | Mean<br>(SE) | SD<br>(SE) | Mean – main effect<br>(SE) | Mean – three-alternative treatment shifter<br>(SE) | SD<br>(SE) | Mean<br>(SE) | SD<br>(SE) | Mean<br>(SE) | SD<br>(SE) |
| Status quo | 4.3028***<br>(0.6083) | 6.9782***<br>(0.5324) | 5.1199***<br>(0.9401) | -0.9279<br>(0.9144) | 6.9978***<br>(0.5220) | 5.8834***<br>(1.9195) | 7.2904***<br>(2.3909) | 5.7004***<br>(0.8861) | 11.0032***<br>(1.4410) |
| Iron 80 µg/l (-50%) | 2.5113***<br>(0.4260) | 1.8553***<br>(0.3489) | 2.5648***<br>(0.8431) | -0.0334<br>(0.8294) | 1.7069***<br>(0.3693) | 5.6059***<br>(2.1168) | 5.4310***<br>(1.8271) | 3.3985***<br>(0.8299) | 4.5739***<br>(0.8180) |
| Iron 40 µg/l (-75%) | 2.8727***<br>(0.4262) | 3.3675***<br>(0.3939) | 3.1221***<br>(0.5743) | 0.0313<br>(0.6147) | 3.2572***<br>(0.3766) | 4.3652**<br>(1.7940) | 5.4945***<br>(1.5515) | 3.4969***<br>(0.8853) | 6.6086***<br>(0.8738) |
| Iron 8 µg/l (-95%) | 3.2586***<br>(0.3234) | 2.6231***<br>(0.2733) | 3.8348***<br>(0.6038) | -0.6509<br>(0.6327) | 2.7451***<br>(0.2948) | 5.9614***<br>(1.7312) | 5.9965***<br>(1.5079) | 4.0400***<br>(0.5561) | 4.6180***<br>(0.5138) |
| Chlorine 0.8 µg/l (-80%) | 2.8154***<br>(0.3450) | 3.2250***<br>(0.2430) | 2.9250***<br>(0.4474) | -0.1712<br>(0.5368) | 3.2806***<br>(0.2379) | 2.1510***<br>(0.6100) | 5.4932***<br>(1.1694) | 2.5991***<br>(0.5973) | 4.3528***<br>(0.4201) |
| Hardness 20°f (-33%) | 4.2252***<br>(0.4424) | 2.6575***<br>(0.2493) | 4.3911***<br>(0.8037) | -0.0015<br>(0.8638) | 2.6261***<br>(0.2426) | 6.6156***<br>(1.8176) | 7.5041***<br>(1.9096) | 4.4679***<br>(0.7944) | 4.9875***<br>(0.6936) |
| Hardness 15°f (-50%) | 5.9022***<br>(0.4081) | 4.1172***<br>(0.2680) | 5.5487***<br>(0.6490) | 0.6647<br>(0.7251) | 3.9599***<br>(0.2572) | 5.9210***<br>(1.0470) | 10.1080***<br>(2.1199) | 6.6968***<br>(0.6900) | 5.8320***<br>(0.5426) |
| *Structural parameters* | | | | | | | | | |
| $\theta$ – scale correction factor for the three-alternative treatment | 0.0773<br>(0.1435) | | - | | | - | | | |
| *Model characteristics* | | | | | | | | | |
| Log-likelihood | -2976.37 | | -2974.91 | | | -2878.37 | | | |
| McFadden pseudo-$R^2$ | 0.4116 | | 0.4119 | | | 0.4310 | | | |
| Ben-Akiva-Lerman pseudo-$R^2$ | 0.6812 | | 0.6808 | | | 0.6877 | | | |
| AIC/n | 0.8061 | | 0.8076 | | | 0.8093 | | | |
| *n* (the number of observations) | 7497 | | 7497 | | | 7497 | | | |
| *k* (the number of parameters) | 45 | | 52 | | | 152 | | | |

*Notes*: ***, **, * represent statistical significance at the 1%, 5% and 10% levels, respectively. SD denotes the standard deviations. Standard errors (SE) are given in parentheses.

Model 1 assumes that the preferences do not differ across the two treatments (except for the scale parameter), while Model 2 relaxes this assumption by including the interactions of each attribute mean with a dummy variable for the three-alternative treatment. All of these interactions are (separately or jointly) insignificant at the 5% significance level, which shows that the number of alternatives does not significantly influence the mean WTP – the elicited preferences do not appear to differ between the two- and three-alternative choice tasks. Using the likelihood ratio test, we formally verify whether imposing the constraint of the equality of the preference parameters is justified. As reported in Table 4 (Model 1 vs. Model 2), the restrictions cannot be rejected at the 5% significance level, i.e., the observed means of the marginal WTP estimates do not differ between the two- and three-alternative choice tasks.

Table 4. Likelihood ratio tests results

| | **Test statistics** | **Degrees of freedom** | **P-value** |
|---|---|---|---|
| Model 1 vs. Model 2 | 2.9017 | 7 | 0.8939 |
| Model 1 vs. Model 3 | 195.9970 | 107 | 0.0000 |
| Model 2 vs. Model 3 | 193.0953 | 100 | 0.0000 |

We verify our findings regarding the equality of the preferences elicited in two- and three-alternative choice tasks by further relaxing the model assumptions. In Model 3, we employ a specification with all treatment-specific preference parameters. The comparison of the mean WTP estimates across the two- and three-alternative data exhibits some differences between the preference parameters; however, these differences are not statistically significant.

As before, we compare Model 1 and Model 3 using the likelihood ratio test to check if the constraint of the equality of the preference parameters for the two- and three-alternative treatments can be rejected. As reported in Table 4, the test result indicates that the model with restrictions (Model 1) should be rejected in favor of the model without restrictions (Model 3). Similarly, when Model 2 and Model 3 are compared, the more flexible Model 3 is preferred. This outcome may look somewhat unexpected given the previous results which indicate no significant differences between the mean marginal WTP estimates across the two- and three-alternative treatments. The reason why Model 3 is statistically preferred lies in the substantial differences in the estimates of the standard deviations of the preference parameters between the two treatments, and their within-treatment correlations. The results of Model 3 show that the standard deviations of the parameters on status quo and on hardness reduction vary considerably between the treatments. However, these differences do not seem to play an important role in the inference of the mean or median WTP of the population. In addition, it should be noted that the variance of preference parameters is typically much less precisely estimated than their mean, especially when the design of the study was not optimized for it, as was the case here. As a result, it may not be surprising that we find differences between the two- and three-alternative estimates of the variances.

In summary, we find that the three-alternative task format does not produce significantly different marginal mean WTP estimates than the two-alternative task format. In other words, people are willing to pay similar amounts for the tap water improvements policies, regardless of whether they state their preferences in two- or three-alternative tasks.

Observing no systematic differences in the mean WTP estimates based on the two- and three-alternative choice data, it may be advantageous to move away from the two-alternative baseline and to use three-alternative choice tasks in terms of higher statistical efficiency (i.e., more information is revealed from a single respondent's choice). To verify the possible efficiency gains, we check whether differences exist in the standard errors of the estimates derived from the two- and three-alternative choices. Table 5 reports the coefficients of variation for each of the parameters, which is the value of the standard error divided by the parameter. It shows that the coefficients of variation are on average 40% lower for the three-alternative treatment than for the two-alternative treatment.

Table 5. Coefficients of variations for the attributes and the alternative-specific constant

| | **Two-alternative treatment** | | **Three-alternative treatment** | |
|---|---|---|---|---|
| | **Coefficient of variation for mean** | **Coefficient of variation for SD** | **Coefficient of variation for mean** | **Coefficient of variation for SD** |
| Status quo | 0.33 | 0.33 | 0.16 | 0.13 |
| Iron 80 µg/l (-50%) | 0.38 | 0.34 | 0.24 | 0.18 |
| Iron 40 µg/l (-75%) | 0.41 | 0.28 | 0.25 | 0.13 |
| Iron 8 µg/l (-95%) | 0.29 | 0.25 | 0.14 | 0.11 |
| Chlorine 0.8 µg/l (-80%) | 0.28 | 0.21 | 0.23 | 0.10 |
| Hardness 20°f (-33%) | 0.27 | 0.25 | 0.18 | 0.14 |
| Hardness 15°f (-50%) | 0.18 | 0.21 | 0.10 | 0.09 |
| Cost | 1.37 | 0.24 | 0.44 | 0.16 |
| Average | 0.44 | 0.26 | 0.22 | 0.13 |

## 5. Discussion and conclusions

We empirically examine whether the number of alternatives per single choice task affects the preferences stated by respondents in a contingent valuation survey. Specifically, we test whether the preferences elicited through two- and three-alternative choice tasks differ significantly. A divergence is expected by the theory because the considered choice task formats may provide different incentives for the respondents answering the survey.

In the public good context of tap water quality improvement, we find no significant differences in the mean WTP estimates obtained from the two- and three-alternative choices. Additionally, our study reveals that the three-alternative choices generally produce more precise WTP estimates, where the

precision is measured by the magnitude of the estimate's standard error corrected for the value of the estimate.

Our finding that the three-alternative task format does not necessarily produce biased estimates provides important evidence for the practical use of the contingent valuation method. In particular, when combined with the increased precision of the welfare measure estimates based on the three-alternative choices and the statistical efficiency gains from a single respondent's choice, the multiple-alternative task formats may be favored over the theoretically suggested, two-alternative task formats.

A question arises about why our results do not mirror the theoretical predictions that when more than two alternatives per choice task are provided, respondents strategically misrepresent their preferences and answer differently than they do in a two-alternative choice task. The existing literature suggests several explanations for these findings.

First of all, strategic manipulation in preference disclosure might appear difficult under task complexity or uncertainty, for example, about other respondents' votes (preferences) or about the voting rule. Task complexity and uncertainty may make it difficult, or even impossible, for a respondent to determine his or her optimal strategic response in a choice task. Conitzer and Sandholm (2002) argue that task complexity can even be intentionally introduced by a researcher as a tool to make strategic manipulation infeasible. With respect to our design, even if the respondents were certain about the preferences of others in the surveyed population, unambiguously determining how the others would vote in each set of provided alternatives might have seemed impossible for them.

Secondly, uncertainty about the preferences of others could also be a reason why we do not observe divergence in the preferences revealed in two- and three-alternative choices. Conitzer *et al.* (2011) show that, when uninformed about others' preferences, a respondent who is potentially voting strategically has no dominant manipulation strategy for the most common voting rules. In our study, the respondents were likely to not know the preferences of others in the surveyed population well; hence, they might have found it difficult to evaluate which policies presented in a choice task had the greatest chance of winning. Tyszler and Schram (2011) emphasize the role of public opinion polls in providing knowledge about others' preferences and observe that such polls may work as a device that allows voters to coordinate on a particular alternative. Such behavior is typically observed, for example, in elections, when polls are available and expectations are well formed. By contrast, in our study of tap water quality improvement, the respondents could not have had enough information about the preferences of others, as no public opinion poll about this issue had been recently conducted.

Thirdly, a yet another factor discouraging respondents from strategically misrepresenting their preferences could be the lack of information about the voting rule that will be used to determine the final decision from the choices in the survey. Being uncertain about the voting rule, respondents might

encounter problems in defining their optimal strategic response (Walsh, 2007; Elkind and Erdélyi, 2012). In our study, the voting rule was not clearly stated; hence, this study characteristic could have also contributed to the lack of observed bias in the three-alternative-based estimates.

In summary, our study shows no statistically significant differences in the value estimates based on two- and three-alternative choice tasks. In addition, the value estimates derived from the three-alternative choices appear to have smaller standard errors than those derived from the two-alternative-based estimates, and respondents' choices are more predictable from the observer's perspective (as indicated by the lower utility function of the error-term variance). Overall, our study contributes to the existing discussion on the use of more than two alternatives per choice task in discrete choice experiments by providing evidence that employing multiple-alternative tasks may increase statistical efficiency without significantly biasing the welfare measures when compared to the welfare measures obtained from data from two-alternative choice tasks. Despite theoretically predicted problems related to the use of more than two choice alternatives, our findings show that respondents do not engage in the strategic misinterpretation of their preferences, possibly because of choice task complexity or uncertainty with respect to other respondents' preferences. Naturally, the obtained results might be sensitive to various characteristics of the valuation study and generalizing our conclusions over the entire contingent valuation domain is discouraged without collecting more evidence and further research.

**References**

Abdalla, C. W., Roach, B. A., and Epp, D. J., 1992. Valuing Environmental Quality Changes Using Averting Expenditures: An Application to Groundwater Contamination. *Land Economics*, 68(2):163-169.

Abrahams, N. A., Hubbell, B. J., and Jordan, J. L., 2000. Joint Production and Averting Expenditure Measures of Willingness to Pay: Do Water Expenditures Really Measure Avoidance Costs? *American Journal of Agricultural Economics*, 82(2):427-437.

Adamowicz, W., Dupont, D., and Krupnick, A. (2006). "Willingness to Pay to Reduce Community Health Risks from Municipal Drinking Water: A Stated Preference Study." In: *3rd World Congress of Environmental and Resource Economics*, Kyoto, Japan.

Adamowicz, W., Dupont, D., Krupnick, A., and Zhang, J., 2011. Valuation of cancer and microbial disease risk reductions in municipal drinking water: An analysis of risk context using multiple valuation methods. *Journal of Environmental Economics and Management*, 61(2):213-226.

Arentze, T., Borgers, A., Timmermans, H., and DelMistro, R., 2003. Transport stated choice responses: effects of task complexity, presentation format and literacy. *Transportation Research Part E: Logistics and Transportation Review*, 39(3):229-244.

Arrow, K., Solow, R., Portney, P. R., Leamer, E. E., Radner, R., and Schuman, H., 1993. Report of the NOAA Panel on Contingent Valuation. *Federal Register*, 58:4601-4614.

Awad, I., and Holländer, R., 2010. Applying Contingent Valuation Method to Measure the Total Economic Value of Domestic Water Services: A Case Study in Ramallah Governorate, Palestine. *European Journal of Economics, Finance and Administrative Sciences*, (20).

Beaumais, O., Briand, A., Millock, K., and Nauges, C., 2014. What are Households Willing to Pay for Better Tap Water Quality? A Cross-Country Valuation Study. FEEM Working Paper No. 24.2014.

Bliemer, M. C. J., Rose, J. M., and Hess, S., 2008. Approximation of Bayesian Efficiency in Experimental Choice Designs. *Journal of Choice Modelling*, 1(1):98-127.

Bresnahan, B. W., Dickie, M., and Shelby, G., 1997. Averting Behavior and Urban Air Pollution. *Land Economics*, 73(3):340-357.

Carlsson, F., Mørkbak, M. R., and 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.

Carson, K. S., Chilton, S. M., and Hutchinson, G. W. (2011). "Bias in Consumer Choices for Public Goods." In: *International Choice Modelling Conference*, Oulton, UK.

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

Carson, R., and Louviere, J., 2011. A Common Nomenclature for Stated Preference Elicitation Approaches. *Environmental and Resource Economics*, 49(4):539-559.

Carson, R. T., and Czajkowski, M., 2014. The Discrete Choice Experiment Approach to Environmental Contingent Valuation. In: *Handbook of choice modelling*, S. Hess and A. Daly, eds., Edward Elgar, Northampton, MA.

Caussade, S., de Dios Ortúzar, J., Rizzi, L. I., and Hensher, D., 2005. Assessing the Influence of Design Dimensions on Stated Choice Experiment Estimates. *Transportation Research Part B: Methodological*, 39(7):621-640.

Central Statistical Office, 2015. Statystyczne Vademecum Samorządowca. Gmina miejska Milanówek. Powiat Grodziski.

Chintakayala, P. K., Hess, S., Rose, J., and Wardman, M. (2009). "Effects of Stated Choice Design Dimensions on Estimates." In: *Inaugural International Choice Modelling Conference*, Harrogate.

Cho, Y., Easter, K. W., McCann, L. M. J., and Homans, F., 2005. Are rural residents willing to pay enough to improve drinking water quality? . *Journal of the American Water Resources Association*, 41(3):729-740.

Collins, A. R., and Steinback, S., 1993. Rural household response to water contamination in west virginia. *Journal of the American Water Resources Association*, 29(2):199-209.

Collins, J. P., and Vossler, C. A., 2009. Incentive compatibility tests of choice experiment value elicitation questions. *Journal of Environmental Economics and Management*, 58(2):226-235.

Conitzer, V., and Sandholm, T., Year. Complexity of manipulating elections with few candidates. Proceedings of the 18th National Conference on Artificial Intelligence (AAAI-02) Edmonton, Canada, 2002, 18: 314-319.

Conitzer, V., Walsh, T., and Xia, L., Year. Dominating Manipulations in Voting with Partial Information. Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, 11: 638-643.

Conlon, B., Dellaert, B. G. C., and van Soest, A. (2001). "Complexity and Accuracy in Consumer Choice: The Double Benefits of Being the Consistently Better Brand." In: *CentER Discussion Paper*, Tillburg University, The Netherlands.

Czajkowski, M., and Budziński, W. (2015). "An insight into the numerical simulation bias – a comparison of efficiency and performance of different types of quasi Monte Carlo simulation methods under a wide range of experimental conditions." In: *Environmental Choice Modelling Conference*, Copenhagen.

Czajkowski, M., Giergiczny, M., and Greene, W. H., 2014. Learning and fatigue effects revisited. Investigating the effects of accounting for unobservable preference and scale heterogeneity. *Land Economics*, 90(2):323-350.

Czajkowski, M., Hanley, N., and LaRiviere, J., 2015. Controlling for the Effects of Information in a Public Goods Discrete Choice Model. *Environmental and Resource Economics*:1-22.

Day, B., Bateman, I. J., Carson, R. T., Dupont, D., Louviere, J. J., Morimoto, S., Scarpa, R., and Wang, P., 2012. Ordering effects and choice set awareness in repeat-response stated preference studies. *Journal of Environmental Economics and Management*, 63(1):73-91.

DeShazo, J. R., and Fermo, G., 2002. Designing Choice Sets for Stated Preference Methods: The Effects of Complexity on Choice Consistency. *Journal of Environmental Economics and Management*, 44(1):123-143.

Dhar, R., and Simonson, I., 2003. The Effect of Forced Choice on Choice. *Journal of Marketing Research*, 40(2):146-160.

Dupont, D. P., and Jahan, N., 2012. Defensive spending on tap water substitutes: the value of reducing perceived health risks. *Journal of Water and Health*, 10(1):56–68.

Elkind, E., and Erdélyi, G., Year. Manipulation under voting rule uncertainty. Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems, International Foundation for Autonomous Agents and Multiagent Systems, 2:627-634.

Genius, M., Hatzaki, E., Kouromichelaki, E. M., Kouvakis, G., Nikiforaki, S., and Tsagarakis, K. P., 2008. Evaluating Consumers' Willingness to Pay for Improved Potable Water Quality and Quantity. *Water Resources Management*, 22(12):1825-1834.

Gibbard, A., 1973. Manipulation of voting schemes: a general result. *Econometrica*, 41(4):587-601.

Gnedenko, E., Gorbunova, Z., and Safonov, G., 2000. Contingent Valuation of Drinking Water Quality in Samara City. EERC Working Paper Series, EERC Research Network, Russia and CIS.

Harrington, W., Krupnick, A. J., and Spofford Jr, W. O., 1989. The economic losses of a waterborne disease outbreak. *Journal of Urban Economics*, 25(1):116-137.

Hensher, D., 2004. Identifying the Influence of Stated Choice Design Dimensionality on Willingness to Pay for Travel Time Savings. *Journal of Transport Economics and Policy*, 38(3):425-446.

Hensher, D., 2006a. How Do Respondents Process Stated Choice Experiments? Attribute Consideration under Varying Information Load. *Journal of Applied Econometrics*, 21(6):861-878.

Hensher, D., 2006b. Revealing Differences in Willingness to Pay due to the Dimensionality of Stated Choice Designs: An Initial Assessment. *Environmental and Resource Economics*, 34(1):7-44.

Janmaat, J., 2007. A Little Knowledge…: Household Water Quality Investment in the Annapolis Valley. *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie*, 55(2):233-253.

Kuksov, D., and Villas-Boas, J. M., 2010. When More Alternatives Lead to Less Choice. *Marketing Science*, 29(3):507-524.

Kwak, S.-J., and Russell, C., 1994. Contingent valuation in Korean environmental planning: A pilot application to the protection of drinking water quality in Seoul. *Environmental and Resource Economics*, 4(5):511-526.

Kwak, S.-Y., Yoo, S.-H., and Kim, C.-S., 2013. Measuring the Willingness to Pay for Tap Water Quality Improvements: Results of a Contingent Valuation Survey in Pusan. *Water*, 5(4):1638-1652.

Laughland, A. S., Musser, L. M., Musser, W. N., and Shortle, J. S., 1993. The opportunity cost of time and averting expenditures for safe drinking water. *Journal of the American Water Resources Association*, 29(2):291-299.

Liu, C.-C., Herriges, J. A., Kling, C. L., Secchi, S., Nassauer, J. I., and Phaneuf, D. J., 2014. A Comparison of Value Elicitation Question Formats in Multiple-Good Contingent Valuation. *Frontiers of Economics in China*, 9(1):85-108.

Luzar, E. J., and Cosse, K. J., 1998. Willingness to pay or intention to pay: The attitude-behavior relationship in contingent valuation. *The Journal of Socio-Economics*, 27(3):427-444.

McFadden, D., 1974. Conditional Logit Analysis of Qualititative Choice Behaviou. In: *Frontiers in Econometrics*, P. Zarembka, ed., Academic Press, New York, 105-142.

Meyerhoff, J., Oehlmann, M., and Weller, P., 2014. The Influence of Design Dimensions on Stated Choices in an Environmental Context. *Environmental and Resource Economics*.

Polyzou, E., Jones, N., Evangelinos, K. I., and Halvadakis, C. P., 2011. Willingness to pay for drinking water quality improvement and the influence of social capital. *The Journal of Socio-Economics*, 40(1):74-80.

Ready, R. C., Whitehead, J. C., and Blomquist, G. C., 1995. Contingent Valuation When Respondents Are Ambivalent. *Journal of Environmental Economics and Management*, 29(2):181–196.

Rolfe, J., and Bennett, J., 2009. The Impact of Offering Two Versus Three Alternatives in Choice Modelling Experiments. *Ecological Economics*, 68(4):1140-1148

Rose, J., Hensher, D., Caussade, S., de Dios Ortúzar, J., and Jou, R.-C., 2009. Identifying Differences in Willingness to Pay due to Dimensionality in Stated Choice Experiments: A Cross Country Analysis. *Journal of Transport Geography*, 17(1):21-29.

Satterthwaite, M. A., 1975. Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10(2):187-217.

Scarpa, R., and Rose, J. M., 2008. Design Efficiency for Non-Market Valuation with Choice Modelling: How to Measure it, What to Report and Why. *Australian Journal of Agricultural and Resource Economics*, 52(3):253-282.

Scarpa, R., Thiene, M., and Hensher, D. A., 2012. Preferences for tap water attributes within couples: An exploration of alternative mixed logit parameterizations. *Water Resources Research*, 48(1):W01520.

Traore, N., Amara, N., and Landry, R., 1999. Households' response to groundwater quality degradation : Results from a household survey in Quebec. *Cahiers d'économie et sociologie rurales*, 52:5-22.

Tyszler, M., and Schram, A., 2011. Information and strategic voting. *Experimental Economics*:1-22.

Um, M.-J., Kwak, S.-J., and Kim, T.-Y., 2002. Estimating Willingness to Pay for Improved Drinking Water Quality Using Averting Behavior Method with Perception Measure. *Environmental and Resource Economics*, 21(3):285-300.

Vossler, C. A., Doyon, M., and Rondeau, D., 2012. Truth in Consequentiality: Theory and Field Evidence on Discrete Choice Experiments. *American Economic Journal: Microeconomics*, 4(4):145-171.

Walsh, T., Year. Uncertainty in preference elicitation and aggregation. Proceedings of the AAAI Conference on Artificial Intelligence, 7:3-8.

Whitehead, J. C., Hoban, T. J., and Houtven, G. V. (1998). "Willingness to Pay and Drinking Water Quality: An Examination of the Averting Behavior Approach." In: *68'th annual conference of the Southern Economic Association*, Baltimore, MD.

Xu, C., Interis, M. G., Petrolia, D. R., and Coatney, K. T. (2013). "Empirical Testing of Strategic Voting and its Implications for Choice Experiments." In: *Southern Agricultural Economics Association Annual Meeting, February 3-5, 2013*, Orlando, Florida.

Yoshida, K. (2009). "Estimating Willingness to Pay for Drinking Water Quality Using Averting Expenditures and Choice Experiments." In: *17th Annual Conference of the European Association of Environmental and Resource Economists*, Amsterdam, The Netherlands

Zawojska, E., and Czajkowski, M., forthcoming. Re-examining empirical evidence on contingent valuation – Importance of incentive compatibility.

Zhang, J., and Adamowicz, W. L., 2011. Unraveling the Choice Format Effect: A Context-Dependent Random Utility Model. *Land Economics*, 87(4):730-743.