Hypothetical Bias Mitigation in Choice Experiments:

Effectiveness of Cheap Talk and Honesty Priming Fade with Repeated Choices

Gregory Howard
Department of Economics, East Carolina University

Brian Roe
Department of AED Economics, Ohio State University

Erik Nisbet
School of Communication, Ohio State University

Jay Martin
Department of Food, Agricultural and Biological Engineering, Ohio State University

Abstract: We design a choice experiment comparing policies that reduce agricultural nutrient pollution and harmful algal blooms in Lake Erie and administer it to Ohio residents using an online survey panel. We compare two treatments that have been found to mitigate hypothetical bias, cheap talk and honesty priming. We find greater sensitivity to price among respondents during choices made immediately following the cheap talk intervention. As additional choices are made, price sensitivity diminishes and eventually matches that of respondents in the control treatment. We find this effect in both our online choice experiments and among respondents to face-to-face choice experiments conducted by de-Magistris, Gracia and Nayga (2013, DNG). Our online implementation of an honesty priming intervention yields no significant change in price sensitivity compared to a control. While DGN (2013) implement an honesty priming intervention that fully mitigates hypothetical bias in a face-to-face setting, we show this effect is also transient, and in later choice exercises we cannot reject the null hypothesis of equality between honesty priming and the control. Our results suggest additional work is required to adapt priming interventions for online settings and to extend the effectiveness of popular hypothetical bias mitigation techniques when respondents face multiple choice tasks.

There is evidence that responses to hypothetical questions and situations may deviate systematically from responses observed in real world interactions. As such, data collected in low- or no-incentive environments may fail to achieve the goal of accurately representing real-world behavior. This deviation, hypothetical bias, affects a wide range of behaviors and preferences. For example, research has found that individuals tend to exhibit greater risk tolerance (Holt and Laury 2005) and greater valuations for a wide range of both private and public goods (Harrison and Rutstrom 2008; List and Gallet 2001) when faced with hypothetical rather than real incentives. This finding poses a particular challenge to the valuation of environmental amenities. While the use value of existing environmental amenities is often estimable using revealed preference methods, in instances where nonuse values may be economically significant or the amenity/program does not currently exist, revealed preference methods are of limited use. When stated preference methods are necessary it is often difficult or impossible to adequately incentivize responses to valuation questions, especially for large-scale projects. For instance, studies that use a referendum format to estimate willingness to pay (WTP) can seldom implement substantive policy based on the outcome of the study referendum.

The difficulty of devising incentive-compatible valuation exercises, coupled with the attractiveness and necessity of using hypothetical questions in some situations, has led researchers to examine the possibility of mitigating or eliminating hypothetical bias when hypothetical techniques are used. This literature has developed and tested several possible remedies, with varying levels of effectiveness. Further, effectiveness may depend on the mode of preference elicitation. Hypothetical questions can be posed via lab experiment, field experiment, phone survey, mail survey, and internet survey. Each mode exerts different levels of researcher control over the respondent’s focus and time commitment. Because of this, it is unclear whether an intervention that eliminates hypothetical bias in one mode will have a similar effect in others. In particular internet surveys, which have grown in popularity in the past decade, may command markedly different levels of respondent time and attention than surveys administered via mail or in a controlled lab setting.

This paper examines data from a survey administered online to Ohio residents that attempts to value a program that reduces agricultural nutrient pollution and harmful algal blooms (HABs) in Lake Erie. We make two unique contributions to the literature on the effect of hypothetical bias mitigation tools influence price sensitivity and WTP estimates. First, we compare the effects of two hypothetical bias treatments in the context of an online survey methodology. One treatment, a "cheap talk" script (Cummings and Taylor 1999), has been shown effective in multiple research contexts, including online surveys (Tonsor and Shupp 2011). The other treatment, honesty priming, has been shown effective in lab experiments (de-Magistris, Garcia and Nayga 2013) but has not yet been tested in an online context.

It is common practice to present each respondent with multiple choice exercises, often as many as sixteen (Louviere, Hensher and Swait 2000). While some research has examined how presenting a greater number of choice sets may influence WTP estimates (Bech, Kjaer and Lauridsen 2011), to our knowledge no study has estimated whether the impact of hypothetical bias mitigation tools varies from early to later choice exercises. Our second contribution is to analyze whether the impact of our interventions is persistent or transient. This is a crucial point of interest that is addressed for the first time in this study. It is possible that an intervention may have as strong impact on response for early exercises that dissipates as time passes and later choices are made.

We find the cheap talk intervention reduces WTP for improved environmental outcomes by about 50% relative to the control, which is in line with estimates of hypothetical bias in the literature. This effect is entirely driven by changes in price sensitivity, specifically increases in the marginal disutility of price; there are no differences in the marginal utility of environmental improvements between the treatments and control. We find no significant difference between the honesty priming treatment and the control. This result is in contrast to the finding of de-Magistris, Garcia and Nayga (2013, hereafter DGN), in which honesty priming reduces WTP in a face-to-face lab setting.

While we find an aggregate effect of the cheap talk treatment, our exercise-level analysis reveals the impact of cheap talk is transient; cheap talk increases price sensitivity relative to the control in the initial choice exercises but not in later choice exercises. This finding is corroborated in a similar exercise-level analysis using data from DGN's in-person lab experiments for both cheap talk and honesty priming treatments and is robust to a variety of model specifications.

The rest of the paper proceeds as follows. Section 1 reviews the relevant literature. Section 2 describes the survey and data. Section 3 outlines the empirical model, Section 4 presents results and Section 5 provides discussion and concludes.

1. Literature

Numerous methods have been developed to eliminate hypothetical bias in choice exercises.[1] The majority of these methods fall under the category of "ex-ante" fixes, meaning they address the problem during the elicitation process.[2] One popular method emphasizes the consequentiality of the questions posed to respondents by convincing them that the questions they face are not purely hypothetical, but may have real consequences (Carson and Groves 2007). This is often achieved by informing respondents that the results of the survey will be used by policy makers to determine future policy. This tactic may curb hypothetical bias, but can also induce strategic responses that do not reflect underlying preferences (Satterhwaite 1975) and at an extreme may require deception when researchers cannot legitimately claim that policy makers will use the survey to inform their decisions.

Another ex-ante approach is the presentation of a cheap talk script prior to the choice exercise. While cheap talk scripts vary in length and specifics, they generally implore the respondent to treat the following hypothetical choice as if it were a real choice involving real money. Longer scripts also educate the respondent on the tendency of survey-takers to overestimate WTP and suggest reasons for this tendency. Several studies have found that cheap talk scripts eliminate hypothetical bias (Cummings and Taylor 1999; Silva *et al.* 2011), mitigate hypothetical bias (Moser, Raffaelli and Notaro 2014; Champ, Moore and Bishop 2009; de-Magistris and Pascucci 2014), or induce differential effects based on respondent attributes and context (List 2001; Silva *et al.* 2012). Other studies document a decrease in WTP but cannot

[1] For a description of competing theories explaining the source of hypothetical bias, see Loomis (2011, 2014).

[2] Another vein of the literature addresses the problem using "ex-post" fixes in which researchers elicit responses that may suffer from hypothetical bias and adjust the analysis to account for this potential bias in one of several ways. Most ex-post fixes use a certainty question following the choice exercise, though the specific use of this question in subsequent analysis varies (Blumenschein *et al.* 2008; Champ, Moore and Bishop 2009; Moore *et al*. 2010). Still another ex-post fix draws not on certainty questions, but instead involves calibrating hypothetical responses using bias-correction factors derived from the literature (Fox *et al*. 1998).

assess the degree of bias mitigation as no fully incentivized treatment is conducted (Whitehead and Cherry 2007; Tonsor and Shupp 2011; Carlsson, Frykblom and Lagerkvist 2005). Blumenschein *et al.* (2008) find no effect from cheap talk, while Morrison and Brown (2009) find that cheap talk overcorrects for hypothetical bias. While the full story is nuanced, it suffices to say that findings for cheap talk are generally positive but mixed. More recently, researchers have utilized an "oath script," which asks respondents to swear an oath to answer truthfully before being presented with the choice exercise. This method has also been effective at mitigating hypothetical bias (Jacquemet *et al.* 2013; Carlsson *et al.* 2013; de-Magistris and Pascucci 2014).

Adapting techniques from social psychology, DGN utilize an "honesty priming" exercise. While the cheap talk and oath scripts overtly encourage respondents to answer honestly, honesty priming engages the respondent in a simple task that endeavors to subconsciously prime subjects for honesty. Several tasks have been demonstrated effective at priming for truthfulness, including the scrambled-sentence method used by DGN (Chartrand and Bargh 1996). Priming for honesty is typically done by engaging the respondent in multiple iterations of a task (like presenting them with five words and asking them to compose a sentence with them). The task primes for honesty when many of the sentences contain words that are related with honesty (truth, forthcoming, sincere, etc.). DGN show that WTP in the honesty priming treatment is significantly less than in the hypothetical baseline treatment and that the honesty priming treatment is not significantly different from the incentivized treatment.

In addition to this work on hypothetical bias, our research also draws on a literature that examines how different survey modes (face-to-face, mail, phone, internet, etc.) may influence respondent behavior. Lindhjem and Navrud (2011) provide both a theoretical examination of survey mode effects and a thorough review of the literature comparing online surveys with other survey modes. The majority of studies find demographic differences between internet survey respondents and respondents from other modes, though the use of internet panels and data weighting techniques can mitigate this issue. The literature comparing WTP between internet and other survey modes is mixed, with some studies failing to reject equality of estimates (Banzhaf *et al*. 2006; Nielsen 2011; Olsen 2009; Windle and Rolfe 2011) and others rejecting equality (Canavari, Nocella and Scarpa 2005; Marta-Pedroso, Freitas and Domingos 2007; Bell, Huber and Viscusi 2011). To our knowledge, only one study has examined the effectiveness of cheap talk scripts in an internet survey setting (Tonsor and Shupp 2011) and no studies have performed a similar examination of honesty priming.

Several studies have considered how early and late choice exercises may differ from each other. Carlsson *et al.* (2012) present 16 choice exercises to subjects, with the last eight being identical to the first eight. They find smaller error variances for the second group of exercises. Further, they find most of the differences that exist between the first and second set of eight choices come from changes in response to the first choice exercise. Ladenburg and Olsen (2008) identify starting point bias in a choice experiment and note that the effect is transient, meaning it fades in later choice exercises. Bateman *et al.* (2008) find that the effects of anchoring dissipate in later choice exercises, as do inconsistencies between single bound and double choice formats. Brouwer *et al.* (2010) administer a choice experiment with certainty follow-up questions after each exercise. They test whether reported certainty changes from early to late exercises and find no evidence of certainty changes when they control for demographics and choice attributes. While each of these studies has explored how respondent behavior may fluctuate over multiple choice exercises, and some studies have affirmed that respondent preferences are not invariant to choice experiment length, no study to this point has considered whether this phenomenon exists in the realm of hypothetical bias mitigation.

2. Survey and Data

We use an online survey with a between-subjects design to compare two treatment tasks (cheap talk and honesty priming) with a neutral control task. In both treatments and the control, these tasks immediately preceded a sequence of five choice exercises. In our priming task, which follows Vinski and Watter (2012), respondents were given six word sets. Each word set contained one target word and three additional words. One of the additional words was a synonym of the target word, while the others additional words had similar meanings without being synonyms. Respondents were asked to enter the synonym in the text space provided. In the honesty priming task, four of the six target words (factual, honest, candid, and sincere) were intended to prime for honesty. In the neutral control respondents completed a similar task for which all six target words did not prime for honesty, but instead were chosen to avoid priming for any specific attitude or behavior (words like gigantic, jolly, and intelligent). This honesty priming task differed from the scrambled sentence task used by DGN, although both tasks have been shown effective in priming honesty and truthful revelation of information (Chartrand and Bargh 1996; Rasinski *et al*. 2004). The cheap talk treatment presented the respondent with a script[3] that resembles other scripts used in the literature. In order to maintain a similar level of

[3] The full cheap talk script presented to respondents: "Later in this survey, you will be presented with a hypothetical choice involving money. No one will actually be paid money based on the decision you make, but you are asked to make the decision as though it would result in the actual payment.
"Studies show that people tend to act differently when they face hypothetical decisions. In other words, they say one thing and do something different. We call this a 'hypothetical bias.' For example, in a recent study, several different groups of people made decisions just like the one you are about to make. Payment was real for one group and

interaction between the control and both treatments, respondents are asked to provide a brief summary of the script in the cheap talk treatment.

The choice experiment asks respondents to rank three options: a status-quo option and two programs aimed at reducing HABs in Lake Erie by reducing nutrient pollution from agriculture. Changes in Lake Erie's watershed during the past three decades reveal tight linkages between lake health and upstream behavior that influences nutrient dynamics. During the 1970s through the mid-1990s, programs targeting point source phosphorus abatement resulted in steady reductions in HABs and improved water clarity (Makarewicz and Bertram 1991; Ludsin et al. 2001). However, since the mid-1990s, Lake Erie has entered a transitional state, due to simultaneous human and environmental influences (Matisoff and Ciborowski 2005). Recently, phosphorus from agricultural runoff has been identified as the dominant source of resurgent HABs (Ohio EPA 2010) and, in the summer of 2014, after the implementation of this survey, HABs caused the municipal water supply for the city of Toledo to become unsafe, requiring more than 400,000 residents to find other sources of water for drinking and bathing for more than two days. Hence, knowledge of Ohio residents' willingness to pay for policies that alter farm nutrient runoff and subsequent HAB-induced losses of ecosystem services is in great demand by policymakers.

hypothetical for the other group, as it will be for you. The results of these studies were that on average, more people expressed a willingness to pay money in the hypothetical group than in the real group.
"How can we get people to think about their decision in a hypothetical situation like they think in a real situation? I think that when we hear about a situation that involves doing something that is basically good, for example helping people in need, improving environmental quality, or anything else, our basic reaction in a hypothetical situation is to think: sure, I would do this. I really would spend the money; I really, really, think I would.
"But when the situation is real, and we would actually have to spend our money, we think a different way. We basically still would like to see good things happen, but when we are faced with the possibility of having to spend money, we think about our options: If I spend money on this, that's money I cannot spend on other things. So, when the payment is real, we act in a way that takes into account the limited amount of money we have. We make the decision while realizing that we just don't have enough money to do everything we might like to do."

Each program has three attributes that are varied: program cost, program effectiveness, and program details. Program cost is the expected annual cost to the respondent's household and took one of four levels ($19, $34, $71, and $102) for each new proposed program. Program effectiveness was captured using three different attributes (fish kills, annual beach closure and water quality advisory days, and a satellite image of the expected annual HAB). New programs took one of three possible levels, which are measured as the percentage reduction of these undesirable outcomes from the baseline status quo outcome (levels are 10%, 20% and 50% reduction).[4] The status-quo number of beach closure and water quality advisory days was determined using 2011-2013 data on algal toxin measurements taken from state park beaches and public water supplies located on Lake Erie. Program details explain the manner in which funds will be collected and the way in which the program will reduce nutrient pollution. This attribute captures two different concepts: the method of program implementation and the avenue through which the program will financially impact households. There are three program implementation methods (voluntary payment for ecosystem service (PES) programs, compulsory regulations, and fertilizer taxes) as well as three ways households could be impacted financially (higher income taxes, higher sales taxes, and higher food prices).

These two concepts could be treated as separate attributes and varied independently of each other, but this approach is problematic since some potential combinations (fertilizer taxes that impact households via higher income taxes and PES programs that impact households via higher food prices, for example) are not credible. Instead, in order to both reduce the dimensionality of the choice experiment and improve the clarity of the program options, we instead chose to integrate the two concepts as a single attribute with five levels (PES programs

[4] In our design, all three indicators of program effectiveness move in unison. This means that a general value of program effectiveness can be identified, but not the value of individual program effectiveness indicators.

funded by income tax, PES programs funded by sales tax, regulation funded by income tax, regulation funded by sales tax, and a fertilizer tax that raises food prices).

The status quo was labeled the "current program." This program was given the same program details (PES funded by sales tax) and effectiveness (90 annual beach closure and water quality advisory days, 5,467 fish die from fish kills annually) for all choice exercises. Additionally subjects were randomly assigned one of two possible program costs for the status quo ($2 or $5). While there was between-subject variation in current program cost, there was no within-subject variation. Figure 1 displays an example choice exercise.

Experimental design was determined using several experimental design macros[5] available in SAS 9.3. Before beginning the design, we developed and applied a restriction macro similar to those found in Kuhfeld (2010) in order to eliminate choice sets with a dominated program.[6] The resulting design of 20 choice exercises achieves a relative D-efficiency of 90%. The design of 20 was converted into four blocks of five choice exercises each using an efficient choice blocking macro. Before launching the survey, an initial choice experiment was pre-tested with approximately 200 respondents using Amazon's Mechanical Turk platform to identify any problems with the credibility of the choice exercises and/or attribute levels. The order of choice exercise presentation was randomized, ensuring that the order of presentation is not confounded with specific attribute levels in our analysis.

The data were collected from a survey of Ohio residents in March 2014. Respondents were recruited using Qualtrics online subject panels. After removing respondents who failed to

[5] They include %mktruns, %mktex, %mktroll, %choiceff, and %mktblock.
[6] Dominated programs were based on program cost and effectiveness, since we made no ex-ante assumptions about the relative desirability of different program detail options.

complete two "focus tests" embedded in the survey,[7] our sample consisted of 1,209 responses. The entire sample, as well as each treatment/control subsample, was representative of the general population of Ohio in several demographic indicators (gender, age and proportion of the population black vs. nonblack), although the sample is skewed toward individuals with more formal education (36% of the sample has completed a four-year college degree, compared with 28% in Ohio). Demographic information is largely identical in each treatment group, shown in Table 1.

3. Empirical Model

We utilize a random utility model. Indirect utility for individual i associated with program j is given by the following equation:

$$U_{ij} = V_{ij} + e_{ij}, \qquad (1)$$

where U_{ij} is latent or unobserved utility, V_{ij} is observable utility, and e_{ij} is the random or unobservable portion utility for each choice. We further specify that utility is a function of a vector of program attributes X_{ij}:

$$U_{ij} = \beta X_{ij} + e_{ij}. \qquad (2)$$

Assuming that errors are i.i.d. and follow a type 1 extreme value distribution, the probability that individual i will select program j as the best is given by

[7] The purpose of these questions is to identify respondents who are not carefully reading and completing the survey. An example of this type of question is as follows: "Sickle cell anemia is simply a different name for malaria. We are checking to see how closely people follow directions. Please select "Not Sure" for this question." Any respondent who does not select "Not Sure" fails the focus test and is removed from the dataset.

$$Pr_{\mathrm{ij}} = \frac{exp(\beta X_{ij})}{\sum_{j=1}^{J} exp(\beta X_{ij})}. \tag{3}$$

This is the standard conditional logit model. Although respondents rank all three choices in our choice experiment, we convert this to a binary choice variable suitable for the conditional logit by assigning 1 to all programs ranked best and 0 to all other programs.[8]

Variations of this standard model that account for individual-level heterogeneity, including random parameters or mixed logit models, latent class models, and attribute non-attendance models, have become increasingly popular in recent years. While these models are useful, and indeed vital, in addressing certain research questions, we contend that our focus on 1) the aggregate impacts of cheap talk and honesty priming and 2) whether and how these aggregate impacts change over the course of a series of made choices does not substantially benefit from the use of heterogeneous preference models. As a robustness check, we estimate the main models in our study using random parameters logit models and find the same qualitative results as those from the conditional logit models presented below.[9]

Table 2 gives a full description of the variables used in our analysis. In our baseline estimation we include only program attributes,

$$U_{ij} = \beta_1 Price_{ij} + \beta_2 Effect_{ij} + \gamma D_{ij} + \beta_3 ProgA_{ij} + \beta_4 ProgB_{ij} + e_{ij}, \tag{4}$$

where $Price_{ij}$ and $Effect_{ij}$ capture program price and effectiveness (where positive values indicate percentage reductions in undesirable outcomes), $ProgA_{ij}$ and $ProgB_{ij}$ are alternative-specific

[8] While the models that follow use the binary choice variable, we additionally run our main model of interest with program rankings as the dependent variable. We use a rank-ordered logit model in this context and find qualitatively similar results, which can be found in the appendix.
[9] Specifically, cheap talk increases price sensitivity relative to the control group, although this effect is transient and disappears in later choices. Honesty priming does not change price sensitivity relative to the control.

constants, and D_{ij} is a series of indicator variables identifying different program detail alternatives.[10] To test the effectiveness of cheap talk (CT) and honesty priming (HP) in influencing willingness to pay relative to the neutral (NP) control, we additionally include the interactions of treatment dummy variables and program[11] attributes:

$$U_{ij} = \beta_1 NP_i{*}Price_{ij} + \alpha_1 CT_i{*}Price_{ij} + \eta_1 HP_i{*}Price_{ij} +$$

$$\beta_2 NP_i{*}Effect_{ij} + \alpha_2 CT_i{*}Effect_{ij} + \eta_2 HP_i{*}Effect_{ij} + \gamma\, NP_i{*}D_{ij} + \qquad (5)$$

$$\rho CT_i{*}D_{ij} + \theta HP_i{*}D_{ij} + \beta_3 ProgA_{ij} + \beta_4 ProgB_{ij} + e_{ij}.$$

Finally, to test whether any impact of our treatments on price sensitivity is persistent or transient, we estimate the following equation:

$$U_{ij} = \sum_{k=1}^{5} \alpha_k CT_i{*}Price_{ij}{*}Ex_k + \sum_{k=1}^{5} \gamma_k HP_i{*}Price_{ij}{*}Ex_k + \sum_{k=1}^{5} \theta_k NP_i{*}Price_{ij}{*}Ex_k$$

$$+ \beta_2 Effect_{ij} + \gamma D_{ij} + \beta_3 ProgA_{ij} + \beta_4 ProgB_{ij} + e_{ij}. \qquad (6)$$

This formulation includes fifteen three-way interactions between price, treatment and exercise (Ex_k). In all models the omitted program details category is a fertilizer tax. All models incorporate robust standard errors that are clustered by respondent.

4. Results

Column I of Table 3 displays the marginal effects of our baseline model. This model, detailed in equation (4), estimates a conditional logit using only program attributes as

[10] Our model assumes changes in program effectiveness have a linear impact on utility. This is not self-evident, so as a robustness check we estimated equation (4) and included a squared term for program effectiveness. The coefficient for squared effectiveness is not significant (p value = 0.133).

[11] We do not interact treatments with alternative-specific constants. Bosworth and Taylor (2012) use such an interaction and find that cheap talk can decrease program participation on the extensive margin as well as on the intensive margin. As a robustness check, we interact treatments with alternative-specific constants and find no significant effect at the 95% level (p values 0.229, 0.852, 0.065 and 0.648).

explanatory variables. Columns II, III and IV add to the baseline model by examining the effect of our two treatments on the price, effect, and details attributes as shown in equation (5). As expected, decreasing program cost and increasing program effectiveness each increase the probability of program selection in all models. While we control for different program detail attributes in our model, we will focus on *Price* and *Effect* for most of this analysis, noting only that subjects tend to prefer PES programs and regulations to fertilizer taxes. A more detailed analysis of our results regarding program details is provided in the reviewer/online appendix. There is also a clear preference for any program relative to the "current situation" status quo, even after controlling for program attribute levels.

Our main finding, which is consistent through all specifications, is the effectiveness of CT and ineffectiveness of HP at increasing price sensitivity relative to the neutral control. We reject the null hypothesis that the coefficient on price in the CT treatment is equal to the coefficient in NP and HP at either the 99% or 95% confidence level in every model, while we uniformly cannot reject the null of equal price effects between HP and NP. We find very few treatment effects when examining non-price attributes like program effectiveness. P values of these comparisons for Column IV of Table 3 are presented in Table 4.

Table 5 presents WTP estimates for increases in program effectiveness for the control and treatment groups when all three sets of program attributes (price, effectiveness, and program details) are allowed to vary by treatment. WTP for an attribute is calculated using the ratio of coefficients between the attribute in question and the price attribute. This is true in the absence of interaction terms. WTP calculations typically become more cumbersome with the inclusion of many interaction terms. However, the interaction terms included in our model allow for simplified WTP calculations. As an example, WTP for a change in program effectiveness under the CT treatment is $\beta_{Effect*CT}/\beta_{Price*CT}$. For *Effect*, our estimate describes WTP for a percentage point increase in program effectiveness (i.e. reducing beach closures/water advisory days and fish kills by 1 percentage point). Estimated WTP in the CT group is half the estimates from HP and the control group. Our disparity between CT and the control group is similar to the gap in WTP found between hypothetical and incentivized groups in much of the literature (List 2001; Murphy *et al.* 2005).

Because WTP estimates are nonlinear combinations of coefficients, traditional tests of differences are inaccurate. To test for differences in our WTP estimates by treatment, we utilize the complete combinatorial method (Poe, Giraud and Loomis 2005), having constructed distributions for each WTP estimate in order to develop confidence intervals using the Krinsky-Robb procedure (Krinsky and Robb 1986; Haab and McConnell 2002). The bottom panel of Table 5 reports *p* values for these tests. We find evidence that estimates of WTP improvements in program effectiveness are significantly lower in the cheap talk treatment than in the control and honesty priming treatments, and find no evidence of differences between the control and honesty priming.

Next, we allow for variations by choice exercise as specified in Equation (6). Table 6 and Figure 2 summarize how price sensitivity changes intertemporally and whether temporal changes in price sensitivity vary by treatment group. This analysis allows us to test whether the effect of our hypothetical bias mitigation techniques persists throughout the entire set of choice exercises or fades over time. Table 6 displays regression results for both a baseline estimation that includes price-exercise interactions (Column I) and an estimation that includes three-way price-treatment-exercise interactions (Column II). Figure 2 displays a comparison of price sensitivity by both treatment and exercise. In corroboration with our previous models, the CT treatment produces greater price sensitivity than the control and HP treatment, and there is no discernible difference between honesty and neutral priming.[12] Importantly, as respondents progress to later exercises, the gap in price sensitivity between CT and the other groups narrows. This suggests that the ability for cheap talk to mitigate hypothetical bias fades during later choice rounds. Indeed, we find the hypothetical bias mitigation effect of cheap talk scripts may decline relatively quickly.

Table 7 adds inference to our findings from Figure 2. We use Wald tests and strongly reject the null of equal price sensitivity between CT and the control group for exercises 1-3, but we can no longer reject the null of equality for exercises 4 and 5. The same pattern holds when CT and HP are compared: we reject the null of equality for exercises 1-3 but not for exercises 4 and 5.

We next apply a similar analysis to WTP, as WTP estimates have greater economic and policy relevance than price sensitivity coefficients. Figure 3 and Table 8 detail changes in WTP estimates by exercise and treatment. We find a similar trend; CT exhibits lower WTP values than either priming group. This gap is larger and statistically significant in early exercises, but the gaps narrow in later exercises and the difference loses statistical significance by the fifth and final exercise. It is worth noting that an outlier WTP estimate of $15.64 is generated for the first exercise of the control treatment. This is due to a price coefficient that is small and not statistically distinguishable from zero, leading to a WTP estimate that is large but contains zero in its 95% confidence interval. We explore addressing this issue using the Carson and Czajkowski method (Carson and Czajkowski 2013), which generates nearly identical WTP estimates but ensures that the distribution of the price coefficient does not include zero. This is

[12] Note that greater levels of price sensitivity correspond to larger (in absolute value) negative numbers, so downward movement along the y-axis in Figure 2 corresponds with greater price sensitivity.

done by taking the negative of price and estimating a mixed logit model that specifies a lognormal distribution for price but constrains the dispersion of the price coefficient to zero. Estimating the models presented in Table 6 using the Carson and Czajkowski method eliminates the WTP outlier and maintains the major result. Specifically, WTP is significantly lower in CT that in HP or NP for early exercises but this significant difference disappears by the fifth exercise.

Robustness Check: DGN Data

It is reasonable to wonder whether our findings are generalizable, or are instead an artifact of some aspect of our research design. To test our conclusions for robustness, we use data from what is, to our knowledge, the only other research comparing cheap talk and honesty priming treatments to a neutral priming control (DGN, 2013). While DGN compare multiple additional treatments (7 total comparison groups), we restrict our focus to the three that most closely resemble the groups in our data.[13]

In principle, an apples-to-apples comparison of the two data sets would be ideal. This would probably entail comparing our data with the first five choice exercises from the DGN data. Unfortunately, two aspects of the DGN design make this comparison problematic. First, the absence of randomization in the DGN data makes it impossible to separate the effect of choice exercise order from the design (specifically, attribute levels) of a specific choice exercise. To deal with this and smooth any exercise-by-exercise variation, we combine exercises in groups of 2 when using DGN data.[14] Second, the DGN choice experiment has an efficient design of 16

[13] We use data from DGN's hypothetical neutral prime, hypothetical honesty prime, and hypothetical cheap talk groups and exclude the hypothetical baseline, real baseline, real neutral prime and real honesty prime groups.
[14] The appendix includes an analysis of all 16 exercises without grouping. Using this model it is clear that price sensitivity varies widely based on the choice exercise, with price sensitivity spiking for several exercises. We also

choice exercises, but any subset of the design, like the first five exercises, is likely to be rather inefficient. If choice exercise order was randomized this would not be an issue, as restricting the observations to the first five each individual encounters would still include all 16 choice exercises. This is not the case for the DGN data, so restricting our analysis to the first five observations will likely create efficiency problems. Indeed, one of the attributes (km2000) appears only once in the first five choice exercises.[15] In light of these issues, we estimate a model using all 16 exercises and estimate price effects by groups of two.

The results of this analysis are detailed in Tables 9 and 10 and Figure 4. In the DGN data, both HP and CT increase price sensitivity. This is contrary to the finding in our data that CT increases price sensitivity while HP does not. When coefficients are allowed to vary by choice exercise, the gap between control and treatment coefficients is larger in early exercises than in later ones, suggesting erosion of the treatment effect. While treatment effects are consistently negative (implying greater price sensitivity in the treatments than the control), Table 8 shows that these differences are more likely to be statistically significant in the early exercises than in later ones. Indeed, the point at which differences between treatments and the control are no longer significant is similar in our data and the DGN data.

While treatment effects dissipate after the early exercises in both datasets, they appear to spike again in the last exercises of the DGN data for the CT treatment. Though we cannot isolate a definitive cause of this phenomenon, one possible explanation for this difference across studies is differences in subject information regarding the number of choice exercises. In our sample,

---

include, as a robustness check, an analysis of the DNG data where exercises are aggregated in groups of four instead of two in the appendix. Our findings are robust to the different aggregations.

[15] We include a model in the appendix that uses only the first five exercises and show that the estimates are nonsensical (for example, the coefficient on price for the fifth exercise is positive for all treatments).

respondents were not told beforehand how many choice exercises they would perform. In DGN, subjects knew they would be making 16 choices. Subjects may choose to act differently for their final few choices, and this tendency may be magnified by the CT treatment. This would explain why such a pattern occurs in the DGN data but not in our own data. This seems plausible, but further research is necessary to adequately test this hypothesis.

Robustness Check: Heterogeneous Preference Models and Variance Issues

To test whether our main results are an artifact of an unrealistic homogeneous preference assumption, we estimate the model from Column II of Table 6 (with treatment-exercise-price interactions) using a mixed logit model (Revelt and Train 1998; McConnell and Tseng 1999). Our model allows for heterogeneity in all non-price variables. This model, whose results are in the appendix, supports our main finding of a transient cheap talk effect on price sensitivity.

Standard conditional and mixed logit models impose a uniform scale parameter, normalized to one. As the scale parameter is proportional to the inverse of the error variance, a homogeneous scale parameter implies homogeneous error variance across all choice situations. Several studies have demonstrated the restrictiveness of this assumption in different contexts (Czajkowski, Giergiczny and Greene 2014; Day *et al.* 2012). This applies to the current study in several respects: we have normalized scale parameters to one, which implies equal variance a) between the status quo and new program options, b) between the cheap talk and priming treatments, and c) between initial and subsequent choice exercises. One or all of these assumptions may be violated, and in this event our coefficient and subsequent WTP estimates may be biased.

As a robustness check, we test whether our mainline results hold in models that allow for scale parameters to be freely estimated. To this end, we estimate three models that are similar to the model from Column II of Table 6 (with treatment-exercise-price interactions) but also allow for heterogeneous error variance by estimating scale parameters. These models allow for heterogeneous error variance by alternative (status quo vs. new program), hypothetical bias treatment (cheap talk vs. honesty prime vs. neutral prime) and choice exercise (first exercise, second exercise, etc.), respectively. Extensive results are available in the appendix.

We find error variance differs by choice, but not by treatment or exercise. Specifically, the status quo option has higher error variance than the new program options. All three scale parameter models support our finding that the effect of cheap talk interventions dissipates in later choice exercises.

5. Discussion and Conclusions

Our exercise-level analysis reveals a conclusion that is novel yet unsurprising: ex-ante treatments that mitigate hypothetical bias in choice experiments have the potential to fade over time. Treatments that are initially very effective did not maintain their potency even though our study used only five exercises, which is modest compared to many designs in the literature. In an attempt to identify whether this result is an artifact of our study or indicative of a more widespread phenomenon, we use data from DGN and find the same pattern for both cheap talk and honesty priming interventions. All treatments that mitigate price sensitivity in the aggregate (cheap talk for our data, both cheap talk and honesty priming for DGN data) exhibit the same basic pattern, regardless of whether they are overt (cheap talk) or inconspicuous (priming) in nature.

This result has important ramifications for future choice experiment design. Choice experiments come in many forms, and a multitude of decisions can impact researchers' ability to accurately elicit preferences. Each decision is not made in a vacuum; instead the appropriateness of one choice depends on other choices. Our analysis suggests that the transient impact of hypothetical bias mitigation techniques may pose problems in a wide number of designs and contexts. Determining how choice experiments can be designed to eliminate this cheap talk and honesty priming erosion is worthy of future consideration.

Our study provides further evidence that cheap talk scripts can significantly mitigate hypothetical bias in online choice exercises and suggests that honesty priming may not be as effective in all choice formats. Although there are many differences between our study and DGN, we believe differences in survey medium (online survey vs. face-to-face lab experiment) are the most likely source of this disparity in honesty priming effectiveness. While the priming tasks are different in the two studies considered here (DGN uses scrambled sentences and our study uses matching synonyms), this is unlikely to be the cause of our different results. Both priming tasks have not only been demonstrated to effectively prime subjects, but have been specifically shown to prime for honesty/truthfulness in other settings.

The repetitions disparity (DGN subjects complete 24 priming tasks while our subjects complete 6) is a more plausible explanation, and indeed we find that subjects spent much more time on the cheap talk task than either priming task in our online survey. It's reasonable to conjecture that our priming intervention was less effective than our cheap talk intervention because of time spent on the task rather than the nature of the task, but the data does not support this. As a test, we control for the interaction of time spent on the treatment/control task with price sensitivity, and further allow for this interaction to vary by treatment group. All interactions of time, treatment and price are not significant, while the general effects persist (i.e. cheap talk still increases sensitivity to price and honesty priming does not).[16] This supports our conclusion that differences in treatment effects are due to the nature of the treatments rather than being a function of time spent on each task.

It is also possible that the nature of the goods being considered (DGN analyzes almonds, a private good, while we analyze pollution reduction, a public good) may be driving the differences we observe, but it is unclear to us why honesty priming should work for private goods but not for public goods. We propose that subjects may be more influenced by a prime when they are more engaged in the priming task, and this engagement may be more easily obtained in a controlled lab setting than in the relatively "hands-off" setting of an online survey. In the context of this study we are unable to rigorously test this proposal, but believe it's a promising avenue for further study.

Importantly, our results suggest that past analyses of choice experiments featuring cheap talk interventions and multiple choice sets might be revisited with exercise-level controls to further explore the robustness of our findings. If our results are supported by further work, interventions that reiterate the main assertions of the cheap talk script at the mid-point of repeated rounds or repeat priming exercises may serve as a "booster shot" for hypothetical bias mitigation. Indeed, such dynamics may explain results found by Ladenburg and Olsen (2014).

[16] *P* values for the price/CT, price/HP, price/CT/Time and price/HP/Time interactions are < 0.005, 0.850, 0.516, and 0.999, respectively.

References

Banzhaf, H., D. Burtraw, D. Evans, and A. Krupnick. 2006. “Valuation of Natural Resource Improvements in the Adirondacks.” *Land Economics,* 82 (3): 445-464.

Bateman, I., D. Burgess, W.G. Hutchinson and D. Matthews. 2008. “Learning Design Contingent Valuation (LDCV): NOAA Guidelines, Preference Learning and Coherent Arbitrariness.” *Journal of Environmental Economics and Management*, 55: 127-141.

Bech, M., T. Kjaer and J. Lauridsen. 2011. “Does the Number of Choice Sets Matter? Results from a Web Survey Applying a Discrete Choice Experiment.” *Health Economics*, 20: 273-286.

Bell, J., J. Huber and W. Viscusi. 2011. “Survey Mode Effects on Valuation of Environmental Goods.” *International Journal of Environmental Research and Public Health*, 8: 1222-1243.

Blumenschein, K., G. Blomquist, M. Johannesson, N. Horn and P. Freemen. 2008. “Eliciting Willingness to Pay without Bias: Evidence from a Field Experiment.” *The Economic Journal*, 118: 114-137.

Bosworth, R and L. Taylor. 2012. “Hypothetical Bias in Choice Experiments: Is Cheap Talk Effective at Eliminating Bias on the Intensive and Extensive Margins of Choice?” *The B.E. Journal of Economic Analysis and Policy*, 12(1): 1-26.

Brouwer, R., T. Dekker, J. Rolfe and J. Windle. 2010. “Choice Certainty and Consistency in Repeated Choice Experiments.” *Environmental and Resource Economics,* 46: 93-109.

Canavari, M., G. Nocella and R. Scarpa. 2005. “Stated Willingness-to-pay for Organic Fruit and Pesticide Ban.” *Journal of Food Products Marketing,* 11 (3): 107-134.

Carlsson, F., M. Kataria, A. Krupnick, E. Lampi, A. Lofgren, P. Qin, and T. Sterner. 2013. “The Truth, the Whole Truth, and Nothing But the Truth – A Multiple Country Test of an Oath Script.” *Journal of Economic Behavior and Organization*, 89: 105-121.

Carlsson, F., P. Frykblom and C. Lagerkvist. 2005. “Using Cheap-talk as a Test of Validity in Choice Experiments.” *Economics Letters,* 89 (2): 147-152.

Carlsson, F., M. Morkbak and S Olsen. 2012. “The First Time is the Hardest: A Test of Ordering Effects in Choice Experiments.” *Journal of Choice Modelling*, 5(2): 19-37.

Carson, R. and M. Czajkowski. 2013. “A New Baseline Model for Estimating Willingness to Pay from Discrete Choice Models.” In: *International Choice Modelling Conference,* Sydney.

Carson, R. and T. Groves. 2007. “Incentive and Informational Properties of Preference Questions.” *Environmental and Resource Economics*, 37: 181-210.

Champ, P., R. Moore and R. Bishop. 2009. “A Comparison of Approaches to Mitigated Hypothetical Bias.” *Agricultural and Resource Economics Review*, 38 (2): 166-180.

Chartrand, T. and J. Bargh. 1996. “Automatic Activation of Impression Formation and Memorization Goals: Nonconscious Goal Priming Reproduces Effects of Explicit Task Instructions.” *Journal of Personality and Social Psychology*, 71 (3): 464-478.

Cummings, R. and L. Taylor. 1999. “Unbiased Value Estimates for Environmental Goods: A Cheap Talk Desing for the Contingent Valuation Method.” *American Economic Review*, 89 (3): 649-665.

Czajkowski, M., M. Giergiczny and W. Green. 2014. “Learning and Fatigue Effects Revisited: Investigating the Effects of Accounting for Unobservable Preference and Scale Heterogeneity.” *Land Economics*, 90 (2): 323-350.

Day, B., I. Bateman, R. Carson, D. Dupont, J. Louviere, S. Morimoto, R. Scarpa and P. Wang. 2012. “Ordering Effects and Choice Set Awareness in Repeated-Response Stated Preference Studies.” *Journal of Environmental Economics and Management*, 63: 73-91.

de-Magistris, T., A. Gracia and R. Nayga. 2013. “On the Use of Honesty Priming Tasks to Mitigate Hypothetical Bias in Choice Experiments.” *American Journal of Agricultural Economics*, 95 (5): 1136-1154.

de-Magistris, T. and S. Pascucci. 2014. “The Effect of the Solemn Oath Script in Hypothetical Choice Experiment Survey: A Pilot Study.” *Economics Letters*, 123: 252-255.

Fox, J., J. Shogren, D. Hayes and J. Kliebenstein. 1998. “CVM-X: Calibrating Contingent Values with Experimental Auction Markets.” *American Journal of Agricultural Economics*, 80 (3): 455-465.

Haab, T. and K. McConnell. 2002. *Valuing Environmental Natural Resources: The Econometrics of Non-Market Valuation*. Northhampton, MA: Edward Elgar Publishing.

Harrison, G. and E. Rutstrom. 2008. “Experimental Evidence on the Existence of Hypothetical Bias in Value Elicitation Methods.” *Handbook of Experimental Economics Results*, 1 (5): 752-767.

Holt, C. and S. Laury. 2005. “Risk Aversion and Incentive Effects: New Data without Order Effects.” *American Economic Review*, 95 (3): 902-904.

Jacquemet, N., R. Joule, S. Luchini and J. Shogren. 2013. “Preference Elicitation under Oath.” *Journal of Environmental Economics and Management*, 65: 110-132.

Krinsky, I. and A. Robb. 1986. “On Approximating the Statistical Properties of Elasticities.” *Review of Economics and Statistics*, 68: 715-9.

Kuhfeld, W. 2010. *Marketing Research Methods in SAS*. Cary, NC: SAS Institute Inc.

Ladenburg, J., and Olsen, S. B. (forthcoming). “Augmenting Short Cheap Talk Scripts with a Repeated Opt-out Reminder in Choice Experiment Surveys.” *Resource and Energy Economics*.

Ladenburg, J. and S. Olsen. 2008. “Gender-specific Starting Point Bias in Choice Experiments: Evidence from an Empirical Study.” *Journal of Environmental Economics and Management*, 56: 275-285.

Lindhjem, H. and S. Navrud. 2011. “Using Internet in Stated Preference Surveys: A Review and Comparison of Survey Modes.” *International Review of Environmental and Resource Economics*, 5: 309-351.

List, J. 2001. “Do Explicit Warnings Eliminate the Hypothetical Bias in Elicitation Procedures? Evidence from Field Auctions for Sports Cards.” *American Economic Review*, 91: 1498-1507.

List, J. and C. Gallet. 2001. “What Experimental Protocol Influence Disparities between Actual and Hypothetical Stated Values?” *Environmental and Resource Economics*, 20: 241-254.

Loomis, J. 2011. “What’s to Know about Hypothetical Bias in Stated Preference Valuation Studies?” *Journal of Economic Surveys*, 25 (2): 363-370.

Loomis, J. 2014. “2013 WAEA Keynote Address: Strategies for Overcoming Hypothetical Bias in Stated Preference Surveys.” *Journal of Agricultural and Resource Economics*, 39 (1): 34-46.

Louviere, J., D. Hensher and J. Swait. 2000. *Stated Choice Methods: Analysis and Applications*. Cambridge: Cambridge University Press.

Ludsin, S.A., Kershner, M.W., Blocksom, K.A., Knight, R.L., and Stein, R.A. 2001. “Life after Death in Lake Erie: Nutrient Controls Drive Fish Species Richness, Rehabilitation,” *Ecological Applications* 11: 731-746.

Makarewicz, J.C., and P. Bertram. 1991. “Evidence for the Restoration of the Lake Erie Ecosystem,” *Bioscience* 41:216-223

Marta-Pedroso, C., H. Freitas and T. Domingos. 2007. “Testing for the Survey Mode Effect on Contingent Valuation Data Quality: A Case Study of Web Based Versus In-person Interviews.” *Ecological Economics*, 62: 388-398.

Matisoff, G., and J. J. H. Ciborowski. 2005. “Lake Erie Trophic Status Collaborative Study,” *Journal of Great Lakes Research* 31(Suppl. 2):1-10.

McConnell, K. and W. Tseng. 1999. “Some preliminary evidence on sampling of alternatives with the random parameters logit.” *Marine Resource Economics*, 14 (4): 317-332.

Moore, R., R. Bishop, B. Provencher and P. Champ. 2010. “Accounting for Respondent Uncertainty to Improve Willingness-to-pay Estimates.” *Canadian Journal of Agricultural Economics*, 58: 381-401.

Morrison, M. and T. Brown. 2009. "Testing the Effectiveness of Certainty Scales, Cheap Talk, and Dissonance-Minimization in Reducing Hypothetical Bias in Contingent Valuation Studies." *Environmental and Resource Economics*, 44: 307-326.

Moser, R., R. Raffaelli and S. Notaro. 2014. "Testing Hypothetical Bias with a Real Choice Experiment using Respondents' Own Money." *European Review of Agricultural Economics*, 41 (1): 25-46.

Murphy, J., P. Allen, T. Stevens and D. Weatherhead. 2005. "A Meta-Analysis of Hypothetical Bias in Stated Preference Valuation." *Environmental and Resource Economics*, 30: 313-325.

Nielson, J. 2011. "Use of the Internet for Willingness-to-pay Surveys: A Comparison of Face-to-face and Web-based Interviews." *Resource and Energy Economics*, 33: 119-129.

Ohio Environmental Protection Agency. 2010. Ohio Lake Erie Phosphorus Task Force Report. Columbus, OH. 109p.

Olsen, S. 2009. "Choosing Between Internet and Mail Survey Modes for Choice Experiment Surveys Considering Non-market Goods." *Environmental and Resource Economics*, 44: 591-610.

Poe, G., K. Giraud and J. Loomis. 2005. "Computational Methods for Measuring the Difference of Empirical Distributions." *American Journal of Agricultural Economics*, 87 (2): 353-365.

Rasinski, K., P. Visser, M. Zagatsky and E. Rickett. 2004. "Using Implicit Goal Priming to Improve the Quality of Self-report Data." *Journal of Experimental Social Psychology*, 41 (3): 321-327.

Revelt, D. and K. Train. 1998. "Mixed Logit with Repeated Choices: Households' Choices of Appliance Efficiency Level." *Review of Economics and Statistics*, 80 (4): 647-657.

Satterhwaite, M. 1975. "Strategy-proofness and Arrow Conditions: Existence and Correspondence Theorems for Voting Procedures and Welfare Functions." *Journal of Economic Theory*, 10: 187-217.

Silva, A., R. Nayga, B. Campbell and J. Park. 2011. "Revisiting Cheap Talk with New Evidence from a Field Experiment." *Journal of Agricultural and Resource Economics*, 36 (2): 280-291.

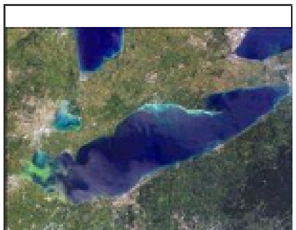
Silva, A., R. Nayga, B. Campbell and J. Park. 2012. "Can Perceived Task Complexity Influence Cheap Talk's Effectiveness in Reducing Hypothetical Bias in Stated Choice Studies?" *Applied Economics Letters*, 19: 1711-1714.

Tonsor, G. and R. Shupp. 2011. "Cheap Talk Scripts and Online Choice Experiments: 'Looking Beyond the Mean'." *American Journal of Agricultural Economics*, 93 (4): 1015-1031.

Vinski, M. and S. Watter. 2012. "Priming Honesty Reduces Subjective Bias in Self-report Measures of Mind Wandering." *Consciousness and Cognition*, 21: 451-455.

Whitehead, J. and T. Cherry. 2007. "Willingness to Pay for a Green Energy Program: A Comparison of Ex-ante and Ex-post Hypothetical Bias Mitigation Approaches." *Resource and Energy Economics*, 29: 247-261.

Windle, J. and J. Rolfe. 2011. "Comparing Responses from Internet and Paper-based Collection Methods in More Complex Stated Preference Environmental Valuation Surveys." *Economic Analysis and Policy*, 41 (1): 83-97.

	Program A	Program B	Current Program
Annual Cost to your Household	$102	$71	$5
Fish Kills	2,733 fish die annually	4,921 fish die annually	5,467 fish die annually
Number of Annual Beach Closure and Water Quality Advisory Days	45 annual beach closure and water quality advisory days	81 annual beach closure and water quality advisory days	90 annual beach closure and water quality advisory days
Annual Expected Size of Lake Erie Algae Bloom (Satellite Photo)			
Program Details	The government will offer Payment for Ecosystem Service (PES) programs. Farmers who choose to enroll in PES programs will be compensated for implementing practices on their farms that reduce nutrient runoff. The program is funded using **<u>state income taxes</u>**. Given your stated income level, the annual cost of this program to your household will be $102	The government will introduce a new tax on fertilizers. This will reduce nutrient use and nutrient runoff in the watershed. The tax will indirectly affect households through higher food prices. Given your stated income level, the annual cost of this program to your household will be $71	The government will offer Payment for Ecosystem Service (PES) programs. Farmers who choose to enroll in PES programs will be compensated for implementing practices on their farms that reduce nutrient runoff. The program is funded using **<u>state sales taxes</u>**. Given your stated income level, the annual cost of this program to your household will be $5

Please provide a ranking for the above programs where 1=Best, 2=Middle, and 3=Worst

	1	2	3
Program A	○	○	○
Program B	○	○	○
Current Program	○	○	○

Figure 1: Choice Exercise Example

Table 1: Balance of Treatments

Variable	Full Sample	Cheap Talk	Honesty Prime	Neutral Prime
Age	45.3	44.9	45.7	45.5
Hispanic	0.02	0.02	0.02	0.02
Female	0.50	0.48	0.50	0.52
White	0.85	0.85	0.86	0.82
Black	0.12	0.11	0.12	0.14
College Grad	0.36	0.36	0.41	0.33
Total Respondents	1210	368	425	417

Table 2: Variable Definitions

Variable	Description
Price	Cost of the program to the respondent's household, in dollars
Effect	Effectiveness of the program, in percentage reduction of negative outcomes
Reg_SalesTax	Dummy equal to 1 if the program uses increased sales taxes to fund increased regulation of farmers
Reg_IncomeTax	Dummy equal to 1 if the program uses increased income taxes to fund increased regulation of farmers
PES_SalesTax	Dummy equal to 1 if the program uses increased sales taxes to fund more voluntary farmer PES programs
PES_IncomeTax	Dummy equal to 1 if the program uses increased income taxes to fund more voluntary farmer PES programs
NP	Dummy equal to 1 if respondent received the neutral priming treatment
CT	Dummy equal to 1 if respondent received the cheap talk treatment
HP	Dummy equal to 1 if respondent received the honesty priming treatment
Ex*N*	Dummy equal to 1 if the choice exercise was the *N*th one faced by the respondent

Table 3: Coefficient Estimates, Models without Exercise Treatments

Variable	I	II	III	IV
Price (Price*NP)	-0.0062** (<0.005)	-0.0045** (<0.005)	-0.0050** (<0.005)	-0.0046** (<0.005)
Price*CT	-	-0.0096** (<0.005)	-0.0095** (<0.005)	-0.0096** (<0.005)
Price*HP	-	-0.0048** (<0.005)	-0.0045** (<0.005)	-0.0047** (<0.005)
Effect (Effect*NP)	0.0259** (<0.005)	0.0259** (<0.005)	0.0269** (<0.005)	0.0266** (<0.005)
Effect*CT	-	-	0.0256** (<0.005)	0.0267** (<0.005)
Effect*HP	-	-	0.0252** (<0.005)	0.0243** (<0.005)
Program A	0.1914* (0.019)	0.1935* (0.018)	0.1936* (0.018)	0.1951* (0.017)
Program B	0.2424** (<0.005)	0.2430** (<0.005)	0.2429** (<0.005)	0.2446** (<0.005)
Program Detail Controls	Yes	Yes	Yes	Yes
Program Detail-Treatment Interactions	No	No	No	Yes

Notes: When two variable names occur in one row, the variable in parentheses is used in Columns II, III and IV. Single and double asterisks (* and **) indicate statistical significance at 95% and 99% confidence levels, respectively. 18,111 observations come from 6,032 choice exercises faced by a total of 1,210 respondents, with robust standard errors clustered by respondent. *P*-values reported in parentheses. The full results of the models presented here are available in the online appendix.

Table 4: P values of Tests for Differences in Coefficent Estimates from Table 3, Column IV

Variable	NP vs. HP	HP vs. CT	NP vs. CT
Price	0.926	**0.012**	**0.010**
Effect	0.457	0.454	0.968

Notes: Bolded values represent comparisons that are statistically different at the 99% confidence level. P values are generated using Wald tests of equality of coefficients.

Table 5: Willingness to Pay Estimates from Table 2, Column IV

	Control	**HP**	**CT**
Effect	**$5.81** [3.70, 14.52]	**$5.13** [3.32, 11.06]	**$2.79** [2.17, 3.87]
P values of Tests for Differences in Estimated WTP Distributions			
	Control vs. HP	HP vs. CT	Control vs. CT
Effect	0.373	**0.013**	**0.006**

Notes: Bolded values indicate significance at 95% confidence level. Numbers in brackets are 95% confidence intervals. Estimates obtained using the Krinsky-Robb procedure with 10,000 draws. Willingness to pay for attribute x is calculated as β_x/β_{price}. In the bottom panel, p values are reported using the complete combinatorial method of testing for differences in distributions (Poe, Giraud and Loomis 2005). A table comparing WTP estimates for all attributes (including program details) is presented in the online appendix.

Table 6: Coefficient Estimates, Models with Exercise Treatments

Variable		I	II
Price (Price*NP)	Ex1	-0.0049** (<0.005)	-0.0017 (0.301)
	Ex2	-0.0072** (<0.005)	-0.0056** (<0.005)
	Ex3	-0.0072** (<0.005)	-0.0049** (<0.005)
	Ex4	-0.0061** (<0.005)	-0.0045* (0.011)
	Ex5	-0.0059** (<0.005)	-0.0063** (<0.005)
Price*CT	Ex1		-0.0091** (<0.005)
	Ex2		-0.0114** (<0.005)
	Ex3		-0.0108** (<0.005)
	Ex4		-0.0090** (<0.005)
	Ex5		-0.0080** (<0.005)
Price*HP	Ex1		-0.0042* (0.012)
	Ex2		-0.0050** (<0.005)
	Ex3		-0.0061** (<0.005)
	Ex4		-0.0052** (<0.005)
	Ex5		-0.0035 (0.055)
Effect, Detail, Program A and Program B Controls?		Yes	Yes

Notes: When two variable names occur in one row, the variable in parentheses is used in Column II. * and ** indicate statistical significance at 95% and 99% confidence levels, respectively. 18,100 observations come from 1,209 respondents and 6,032 choice exercises, with robust standard errors clustered by respondent. *P*-values reported in parentheses.

Figure 2: Price Coefficients by Treatment and Exercise

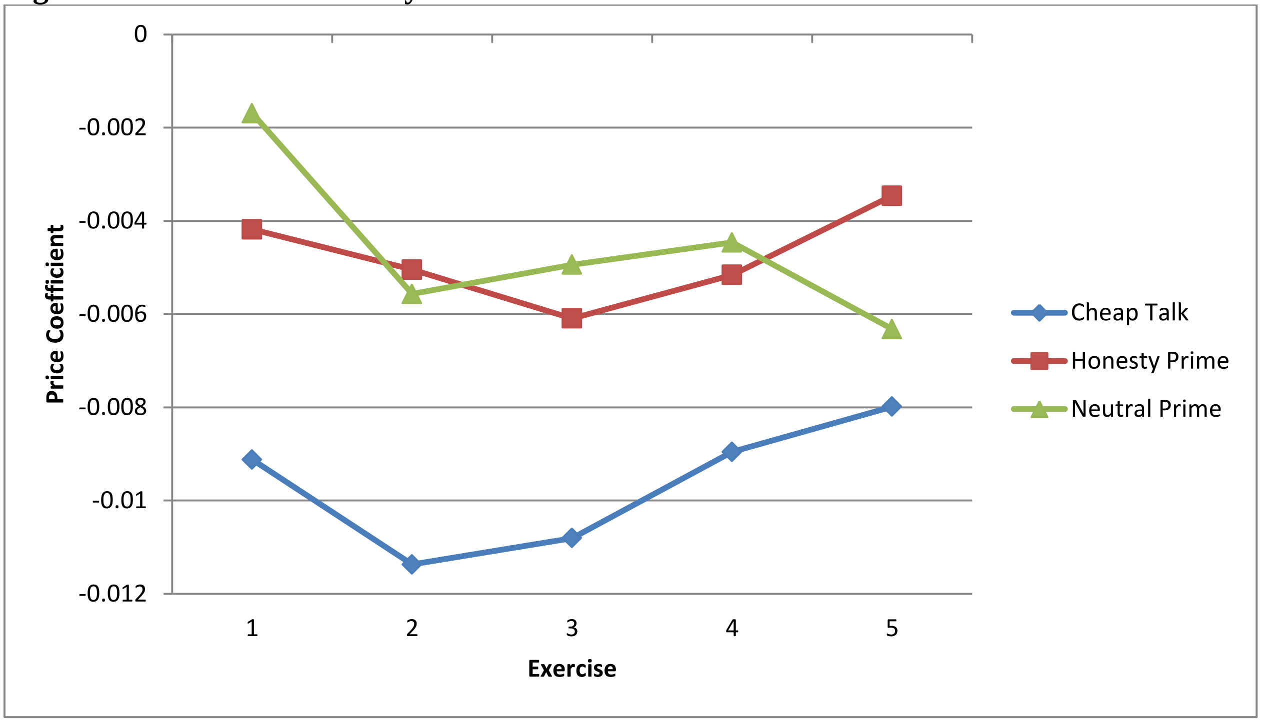
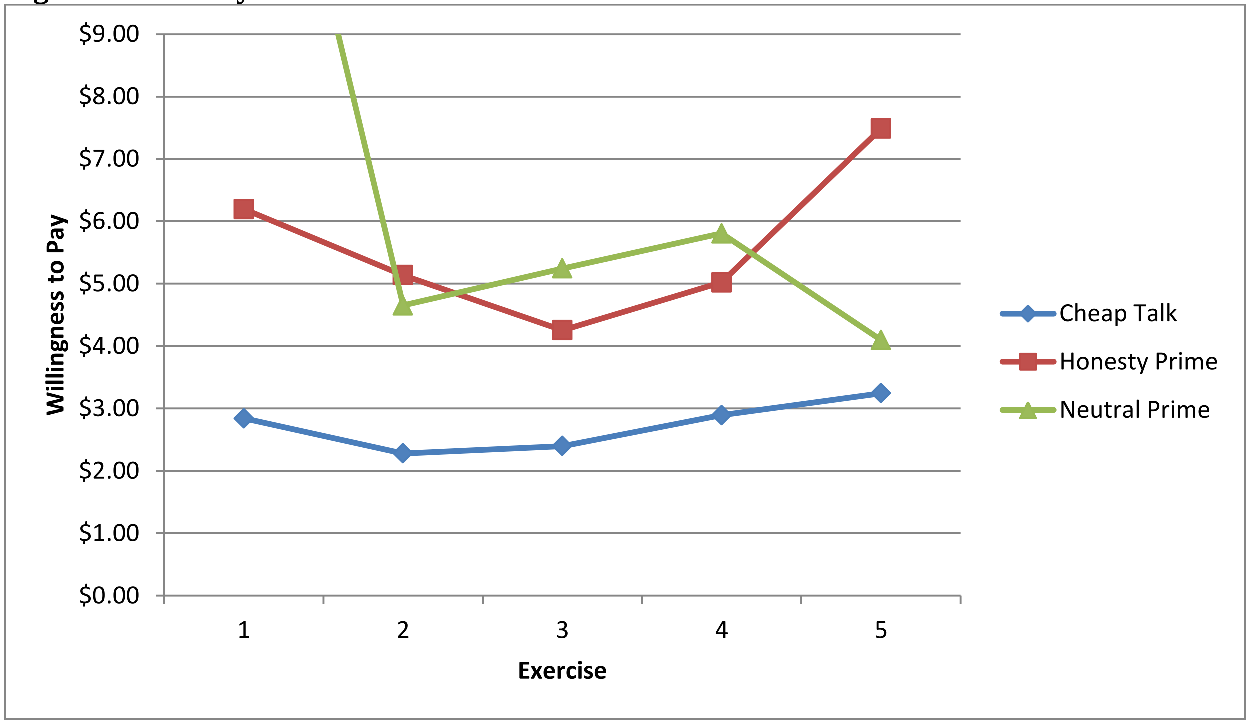


Table 7: Tests for Equality of Price Coefficient Estimates: P Values

	Ex1	Ex2	Ex3	Ex4	Ex5
CT=HP	**0.024**	**< 0.005**	**0.039**	0.106	0.053
CT=Control	**< 0.005**	**0.011**	**0.010**	0.056	0.470
HP=Control	0.240	0.811	0.605	0.758	0.217

Notes: Bolded values indicate significance at 95% confidence level. The null hypothesis is equality of the price coefficient for both groups. Comparisons of treatments use Wald Tests.

Figure 3: WTP by Treatment and Exercise



Notes: WTP for the first exercise neutral priming treatment is $15.31. This outlier value is due to a price coefficient that is close to zero.

Table 8: Tests for Equality of WTP Estimates: P Values

	Ex1	Ex2	Ex3	Ex4	Ex5
CT=HP	**0.023**	**0.005**	**0.023**	0.058	0.061
CT=Control	0.157	**0.008**	**0.009**	**0.037**	0.246
HP=Control	0.288	0.413	0.314	0.389	0.155

Notes: Bolded values indicate significance at 95% confidence level. The null hypothesis is equality of the WTP estimate for both groups. Distributions of each WTP are estimated using the Krinsky-Robb procedure and p values are reported using the complete combinatorial method of testing for differences in distributions (Poe, Giraud and Loomis 2005).

Table 9: Differential Treatment Effects by Choice Exercise using data from DGN

Variable		Coefficient
Price	**Ex1-2**	-1.2126** (<0.005)
	Ex3-4	-1.2263** (<0.005)
	Ex5-6	-1.5030** (<0.005)
	Ex7-8	-1.4589** (<0.005)
	Ex9-10	-1.4049** (<0.005)
	Ex11-12	-1.4022** (<0.005)
	Ex13-14	-1.6041** (<0.005)
	Ex15-16	-1.4452** (<0.005)
Price*CT	**Ex1-2**	-0.4002 (0.059)
	Ex3-4	-0.4066* (0.013)
	Ex5-6	-0.2125 (0.231)
	Ex7-8	-0.1827 (0.362)
	Ex9-10	-0.1835 (0.236)
	Ex11-12	-0.1891 (0.577)
	Ex13-14	-0.1969 (0.214)
	Ex15-16	-0.4283* (0.035)

Notes: * and ** indicate statistical significance at 95% and 99% confidence levels, respectively. 7,632 observations come from 159 respondents, with robust standard errors clustered by respondent.

Table 9 Continued

Price*HP	**Ex1-2**	-0.4283* (0.019)
	Ex3-4	-0.3432* (0.036)
	Ex5-6	-0.3060 (0.060)
	Ex7-8	-0.2931 (0.120)
	Ex9-10	-0.1320 (0.373)
	Ex11-12	-0.2815 (0.432)
	Ex13-14	-0.2677 (0.093)
	Ex15-16	-0.2646 (0.183)
Controls for other Attributes and “No Buy” Option		Yes

Notes: * and ** indicate statistical significance at 95% and 99% confidence levels, respectively. 7,632 observations come from 159 respondents, with robust standard errors clustered by respondent. *P*-values reported in parentheses.

Figure 4: Price Coefficients by Treatment and Exercise using data from DGN

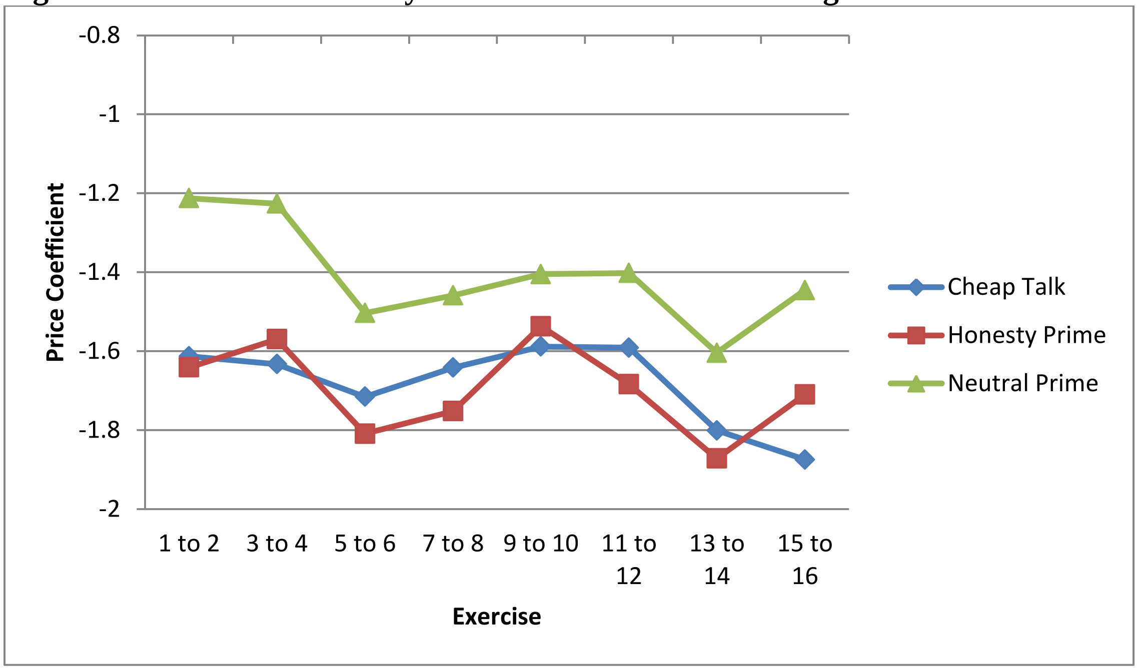


Table 10: Tests of Equality P Values using data from DGN

	Ex1-2	Ex3-4	Ex5-6	Ex7-8	Ex9-10	Ex11-12	Ex13-14	Ex15-16
CT=HP	0.884	0.645	0.585	0.561	0.728	0.756	0.647	0.382
CT=Control	0.059	0.013*	0.231	0.362	0.236	0.577	0.214	0.035*
HP=Control	0.019*	0.036*	0.060	0.120	0.373	0.432	0.093	0.183

Notes: * and ** indicate statistical significance at 95% and 99% confidence levels, respectively. The null hypothesis is equality of the price coefficient for both groups. Comparisons of treatments with the control use t-tests, while comparisons of the two treatments use Wald Tests.

Reviewer Appendix

Table A1: Coefficient Estimates, Models without Exercise Treatments

Variable	I	II	III	IV
Price (Price*NP)	-0.0062** (<0.005)	-0.0045** (<0.005)	-0.0050** (<0.005)	-0.0046** (<0.005)
Price*CT	-	-0.0096** (<0.005)	-0.0094** (<0.005)	-0.0095** (<0.005)
Price*HP	-	-0.0047** (<0.005)	-0.0044** (<0.005)	-0.0047** (<0.005)
Effect (Effect*NP)	0.0258** (<0.005)	0.0258** (<0.005)	0.0269** (<0.005)	0.0266** (<0.005)
Effect*CT	-	-	0.0254** (<0.005)	0.0265** (<0.005)
Effect*HP	-	-	0.0253** (<0.005)	0.0243** (<0.005)
Reg_SalesTax (Reg_SalesTax*NP)	0.1423* (0.029)	0.1427* (0.030)	0.1430* (0.030)	-0.0352 (0.760)
Reg_SalesTax*CT	-	-	-	0.1497 (0.197)
Reg_SalesTax*HP	-	-	-	0.3064** (0.006)
Reg_IncomeTax (Reg_IncomeTax*NP)	0.2115** (<0.005)	0.2136** (<0.005)	0.2134** (<0.005)	0.1690 (0.156)
Reg_IncomeTax*CT	-	-	-	0.1557 (0.198)
Reg_IncomeTax*HP	-	-	-	0.3090** (0.008)
PES_SalesTax (PES_SalesTax*NP)	0.2907** (<0.005)	0.2923** (<0.005)	0.2924** (<0.005)	0.2432** (0.009)
PES_SalesTax*CT	-	-	-	0.3596** (<0.005)
PES_SalesTax*HP	-	-	-	0.2835** (<0.005)
PES_IncomeTax (PES_IncomeTax*NP)	0.1110 (0.161)	0.1054 (0.185)	0.1056 (0.184)	-0.0302 (0.820)
PES_IncomeTax*CT	-	-	-	0.0946 (0.482)
PES_IncomeTax*HP	-	-	-	0.2504* (0.047)
Program A	0.1912* (0.019)	0.1932* (0.018)	0.1934* (0.018)	0.1949* (0.017)
Program B	0.2420** (<0.005)	0.2425** (<0.005)	0.2425** (<0.005)	0.2442** (<0.005)

Notes: When two variable names occur in one row, the variable in parentheses is used in Columns II, III and IV. Single and double asterisks (* and **) indicate statistical significance at 95% and 99% confidence levels, respectively. 18,100 observations come from 6,032 choice exercises faced by a total of 1,209 respondents, with robust standard errors clustered by respondent. *P*-values reported in parentheses.

Table A2: Willingness to Pay Estimates from Table 2, Column IV

	Control	**HP**	**CT**
Effect	**$5.82** [3.70, 14.52]	**$5.13** [3.32, 11.06]	**$2.79** [2.17, 3.87]
Reg*SalesTax	-$7.70 [-81.16, 46.70]	**$64.56** [18.36, 158.74]	$15.80 [-8.60, 40.91]
Reg*IncomeTax	$36.96 [-12.44, 162.40]	**$65.10** [14.70, 194.64]	$16.44 [-8.27, 47.98]
PES*SalesTax	**$53.19** [10.63, 176.74]	**$59.72** [20.02, 154.85]	**$37.96** [18.30, 66.31]
PES*IncomeTax	-$6.61 [-81.66, 64.80]	$52.76 [-0.12, 150.25]	$9.98 [-19.37, 39.42]
P values of Tests for Differences in Estimated WTP Distributions			
	Control vs. HP	HP vs. CT	Control vs. CT
Effect	0.373	**0.014**	**0.006**
Reg*SalesTax	**0.025**	**0.036**	0.201
Reg*IncomeTax	0.277	0.061	0.263
PES*SalesTax	0.437	0.207	0.293
PES*IncomeTax	0.081	0.084	0.308

Notes: Bolded values indicate significance at 95% confidence level. Numbers in brackets are 95% confidence intervals. Estimates obtained using the Krinsky-Robb procedure with 10,000 draws. Willingness to pay for attribute x is calculated as β_x/β_{price}. In the bottom panel, p values are reported using the complete combinatorial method of testing for differences in distributions (Poe, Giraud and Loomis 2005).

Figure A1: Price Coefficients by Treatment and Exercise using data from de-Magistris, Gracia and Nayga (2013); No Grouping

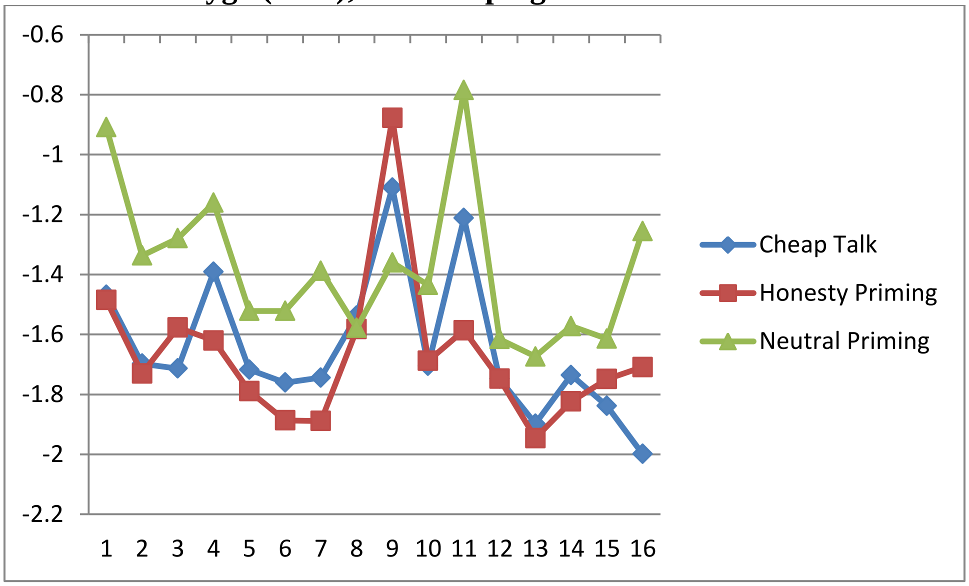


Table A3: Differential Treatment Effects by Choice Exercise using data from de-Magistris, Gracia and Nayga (2013); Groups of Four

Variable		Coefficient
Price	**Ex1-4**	-1.2147** (<0.005)
	Ex5-8	-1.4795** (<0.005)
	Ex9-12	-1.4052** (<0.005)
	Ex13-16	-1.5351** (<0.005)
Price*CT	**Ex1-4**	-0.4037** (0.007)
	Ex5-8	-0.1990 (0.246)
	Ex9-12	-0.1834 (0.238)
	Ex13-16	-0.2890 (0.055)
Price*HP	**Ex1-4**	-0.3689** (0.010)
	Ex5-8	-0.3003* (0.048)
	Ex9-12	-0.1542 (0.309)
	Ex13-16	-0.2702 (0.051)
Controls for other Attributes and "No Buy" Option		Yes

Notes: * and ** indicate statistical significance at 95% and 99% confidence levels, respectively. 7,632 observations come from 159 respondents, with robust standard errors clustered by respondent. *P*-values reported in parentheses.

Table A4: Tests of Equality P Values using data from de-Magistris, Gracia and Nayga (2013); Groups of Four

	Ex1-4	Ex5-8	Ex9-12	Ex13-16
CT=HP	0.786	0.522	0.839	0.888
CT=Control	0.007**	0.246	0.238	0.055
HP=Control	0.010**	0.048*	0.309	0.051

Notes: * and ** indicate statistical significance at 95% and 99% confidence levels, respectively. The null hypothesis is equality of the price coefficient for both groups. Comparisons of treatments with the control use *t*-tests, while comparisons of the two treatments use Wald Tests.

Figure A2: Price Coefficients by Treatment and Exercise using data from de-Magistris, Gracia and Nayga (2013); Groups of Four

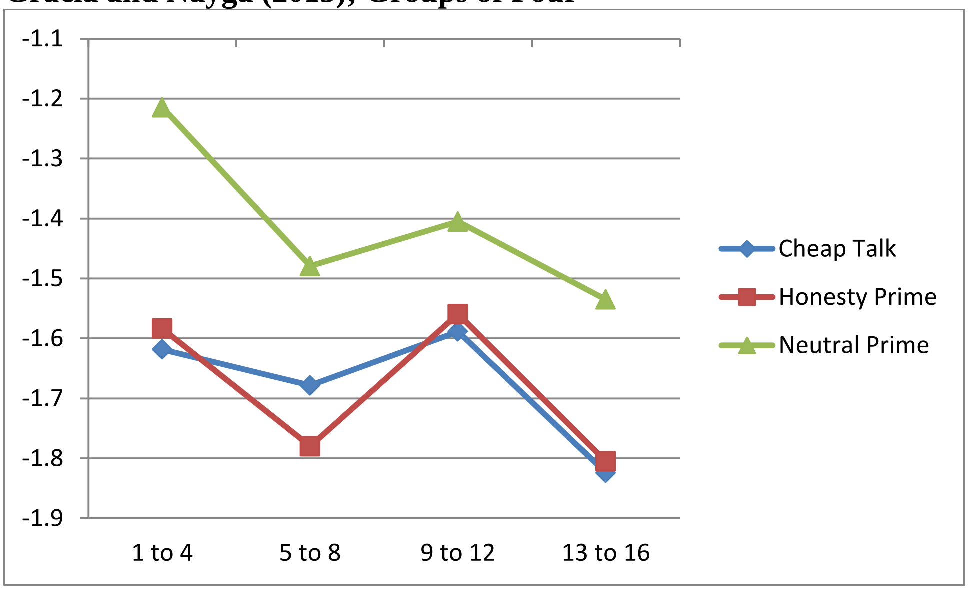


Table A5: Differential Treatment Effects by Choice Exercise using data from de-Magistris, Gracia and Nayga (2013); No Groups, First Five Exercises Only

Variable		Coefficient
Price*NP	**Ex1**	-3.4017* (0.017)
	Ex2	-1.2574** (<0.005)
	Ex3	-3.2190** (<0.005)
	Ex4	-1.3018** (<0.005)
	Ex5	1.6619 (0.294)
Price*CT	**Ex1**	-4.2476** (<0.005)
	Ex2	-1.6759** (<0.005)
	Ex3	-3.6617** (<0.005)
	Ex4	-1.4751 (<0.005)
	Ex5	1.4501 (0.366)
Price*HP	**Ex1**	-4.2720** (<0.005)
	Ex2	-1.7123** (<0.005)
	Ex3	-3.5230** (<0.005)
	Ex4	-1.6448** (<0.005)
	Ex5	1.3737 (0.385)
Controls for other Attributes and "No Buy" Option		Yes

Notes: * and ** indicate statistical significance at 95% and 99% confidence levels, respectively. 7,632 observations come from 159 respondents, with robust standard errors clustered by respondent. *P*-values reported in parentheses.

Figure A3: Price Coefficients by Treatment and Exercise, Mixed Logit

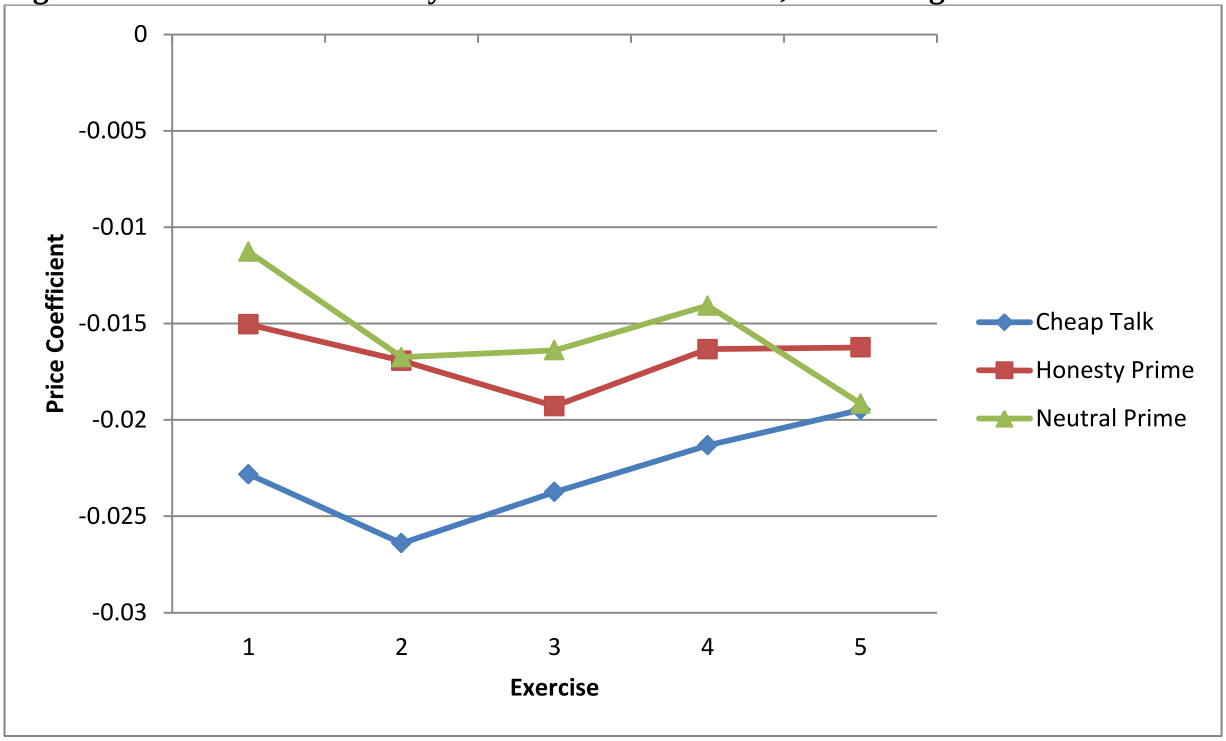


Table A6: Tests for Equality of Price Coefficient Estimates, Mixed Logit: P Values

	Ex1	Ex2	Ex3	Ex4	Ex5
CT=HP	0.0785	**0.024**	0.270	0.232	0.462
CT=Control	**0.010**	**0.027**	0.083	0.081	0.943
HP=Control	0.367	0.964	0.453	0.575	0.491

Notes: Bolded values indicate significance at 95% confidence level. The null hypothesis is equality of the price coefficient for both groups. Comparisons of treatments use Wald Tests.

Figure A4: Price Coefficients by Treatment and Exercise, Rank-Ordered Logit

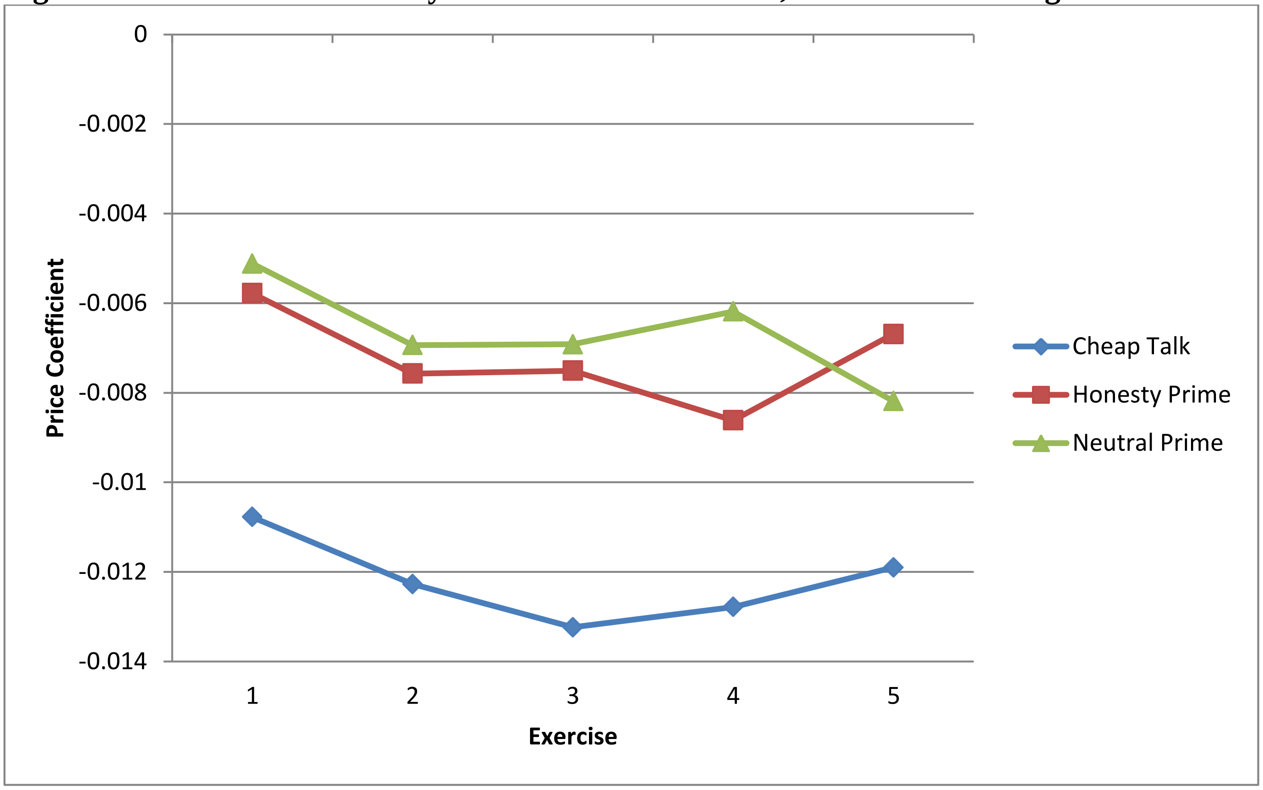


Table A7: Tests for Equality of Price Coefficient Estimates using Rank-Ordered Logit: P Values

	Ex1	**Ex2**	**Ex3**	**Ex4**	**Ex5**
CT=HP	**0.015**	**0.026**	**0.007**	0.067	**0.020**
CT=Control	**0.007**	**0.013**	**< 0.005**	**< 0.005**	0.090
HP=Control	0.748	0.761	0.778	0.268	0.497

Notes: Bolded values indicate significance at 95% confidence level. The null hypothesis is equality of the price coefficient for both groups. Comparisons of treatments use Wald Tests.

Figure A5: Price Coefficients by Treatment and Exercise, Scale Parameters Vary by Choice

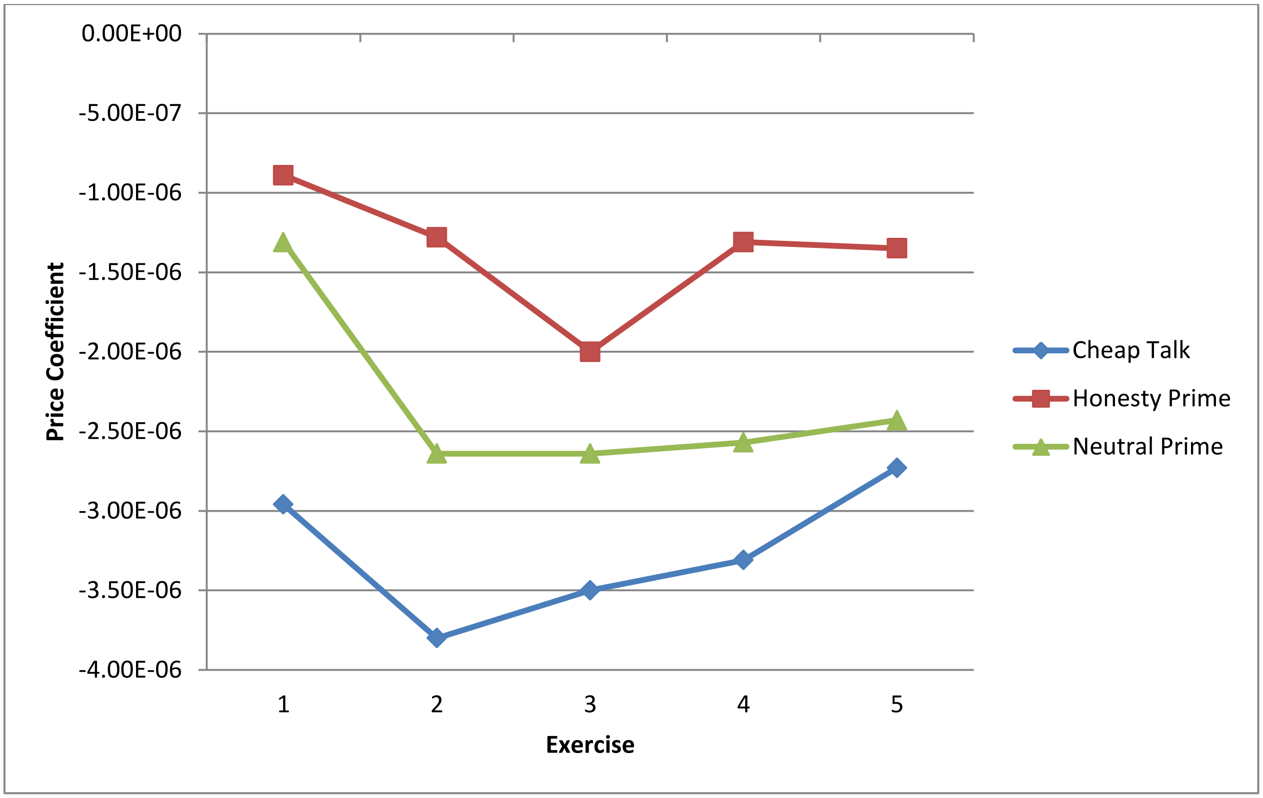


Table A8: Tests for Equality of Price Coefficient Estimates Scale Parameters Vary by Choice: P Values

	Ex1	Ex2	Ex3	Ex4	Ex5
CT=HP	**< 0.005**	**< 0.005**	**0.030**	**< 0.005**	0.050
CT=Control	**0.017**	0.093	0.219	0.302	0.674
HP=Control	0.527	**0.046**	0.350	0.067	0.118

Notes: Bolded values indicate significance at 95% confidence level. The null hypothesis is equality of the price coefficient for both groups. Comparisons of treatments use Wald Tests.

Figure A6: Price Coefficients by Treatment and Exercise, Scale Parameters Vary by Treatment

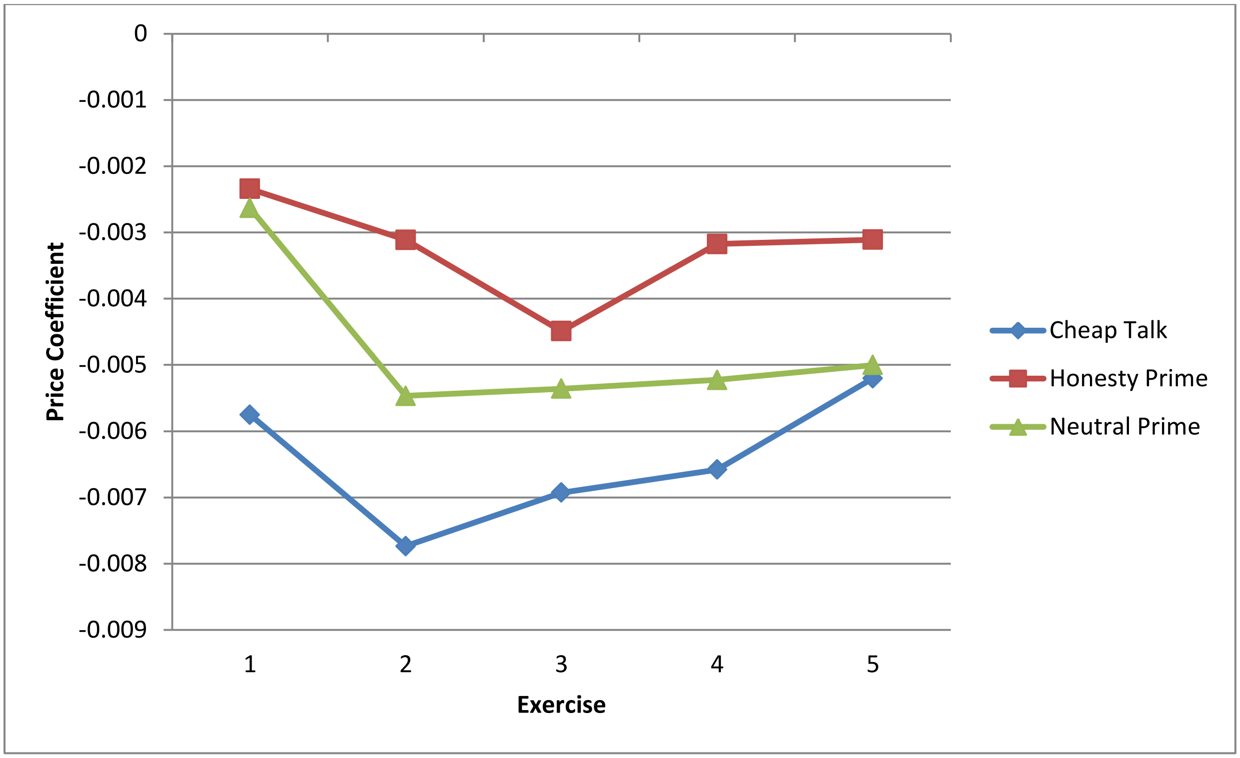


Table A9: Tests for Equality of Price Coefficient Estimates Scale Parameters Vary by Treatment: P Values

	Ex1	**Ex2**	**Ex3**	**Ex4**	**Ex5**
CT=HP	**0.023**	**< 0.005**	0.103	**0.024**	0.165
CT=Control	**0.039**	0.137	0.306	0.386	0.897
HP=Control	0.837	0.089	0.533	0.142	0.180

Notes: Bolded values indicate significance at 95% confidence level. The null hypothesis is equality of the price coefficient for both groups. Comparisons of treatments use Wald Tests.

Figure A7: Price Coefficients by Treatment and Exercise, Scale Parameters Vary by Exercise

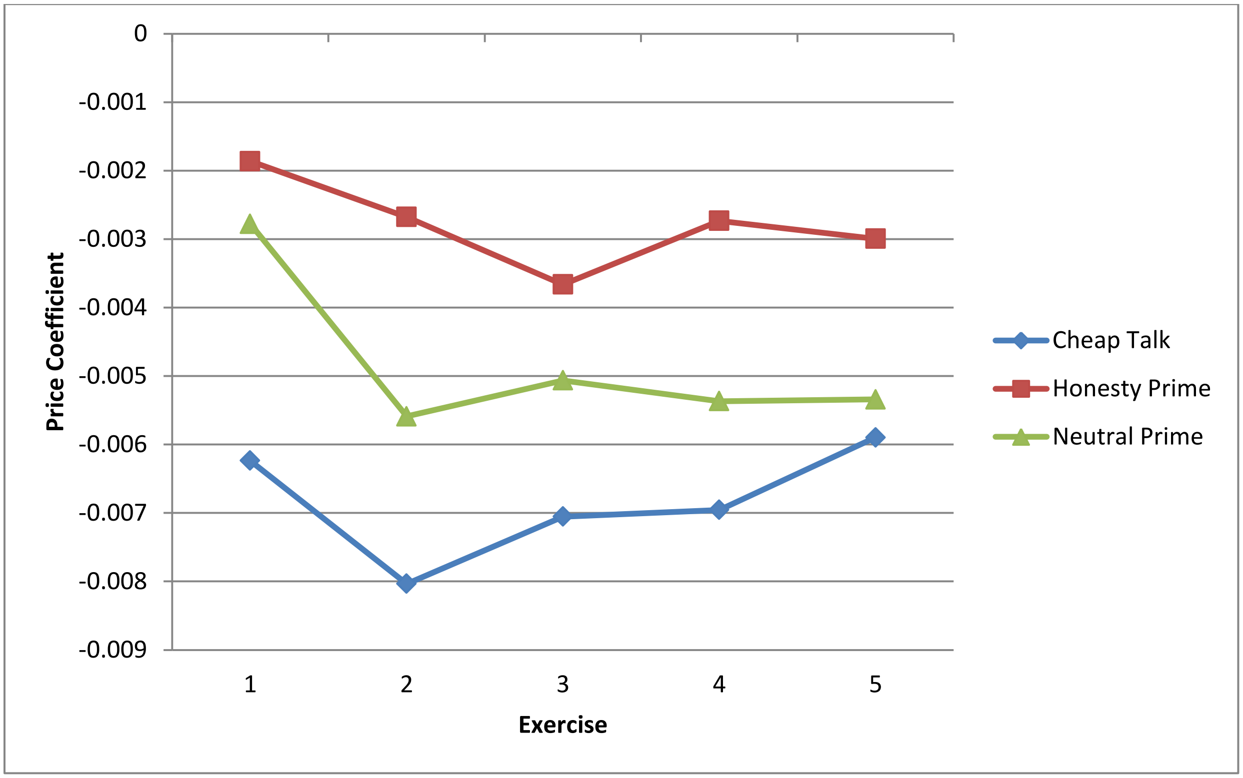


Table A10: Tests for Equality of Price Coefficient Estimates Scale Parameters Vary by Exercise: P Values

	Ex1	**Ex2**	**Ex3**	**Ex4**	**Ex5**
CT=HP	**< 0.005**	**< 0.005**	**0.047**	**0.009**	0.053
CT=Control	**0.021**	0.113	0.233	0.307	0.701
HP=Control	0.528	0.052	0.390	0.084	0.109

Notes: Bolded values indicate significance at 95% confidence level. The null hypothesis is equality of the price coefficient for both groups. Comparisons of treatments use Wald Tests.