On the Use of Virtual Reality in Mitigating Hypothetical Bias in Choice Experiments

Di Fang, Rodolfo M. Nayga Jr, Grant H. West, Claudia Bazzani, Wei Yang, Benjamin C. Lok, Charles E. Levy, and Heather A. Snell

Choice experiments have been extensively used in many stated preference studies and disciplines. In order to give external validity to stated preferences, participants must behave the same in an experiment setting as they do in reality. The difference between a decision made in a real experiment and a hypothetical experiment is commonly attributed to "hypothetical bias." In this paper, we investigate whether virtual reality can reduce hypothetical bias in choice experiments. We conducted a set of experiments using the Food and Drug Administration's new Nutrition Facts Label for yogurts. Participants were randomly assigned to non-hypothetical and hypothetical experiments, which differed in terms of visual presentations of the product in question: text, picture, and a virtual grocery store. Given the absence of real markets for the new label at the time of the research, we approximate the real market with a binding choice in our non-hypothetical experiments. Our results suggest that although the hypothetical bias with virtual reality is not statistically significantly different from those using text or picture, it is the smallest. Among participants who did not exhibit high simulator discomfort, results suggest that virtual reality can significantly reduce hypothetical bias in choice experiments.

Key words: Choice experiment, hypothetical bias, virtual reality.

JEL codes: C91, D9, I1.

Choice experiments (CE) are arguably one of the most popular stated preference elicitation methods being used by applied economists in valuing private and public goods. CE is now extensively used in marketing, agricultural economics, transportation, health, and environmental economics. Willingness to pay (WTP) values and preferences from CEs are used to inform policy, program, or business decisions. The advantage of a CE when evaluating the potential of new products or policies is that it allows the identification of preferences through marginal substitution between attributes. The downside, shared by most stated preference methods, is the real possibility of hypothetical bias, which is well documented in the literature (List and Gallet 2001; Lusk and Schroeder 2004; Chowdhury et al. 2011; De-Magistris, Gracia, and Nayga

---

Di Fang is an assistant professor at the Department of Agricultural Economics and Agribusiness at the University of Arkansas. Rodolfo M. Nayga Jr is a distinguished professor and Tyson Endowed Chair in Food Policy Economics, Department of Agricultural Economics and Agribusiness at the University of Arkansas. Grant H. West is a research program associate at the Department of Agricultural Economics and Agribusiness at the University of Arkansas. Claudia Bazzani is an assistant professor at the Department of Business Administration at the University of Verona. Wei Yang is a research program associate at the Department of Agricultural Economics and Agribusiness at the University of Arkansas. Benjamin C. Lok is a professor at the Computer and Information Sciences and Engineering Department at the University of Florida. Charles E. Levy is the chief of the Physical Medicine and Rehabilitation Service at the North Florida/South Georgia Veterans Health System at the Center for Arts in Medicine and Department of Occupational Therapy at the University of Florida. Heather A. Snell is a research program associate at the Department of Agricultural Economics and Agribusiness at the University of Arkansas.
The contents do not represent the views of the U.S. Department of Veterans Affairs or the United States government.
Correspondence to be sent to: difang@uark.edu

Amer. J. Agr. Econ. 103(1): 142–161; doi:10.1111/ajae.12118
Published online July 19, 2020

2013; Bello and Abdulai 2016; Penn and Hu 2018).

Hypothetical bias is a concept first introduced by Schulze, d'Arge, and Brookshire (1981) as "the potential error induced by not confronting the individual with an actual situation." It refers to the difference in valuations made by individuals in hypothetical settings as opposed to those made in real situations. As valuations are often measured in terms of WTP, many studies have found that individuals overstate their WTP in hypothetical situations (List and Gallet 2001; Murphy et al. 2005; Chowdhury et al. 2011; De-Magistris, Gracia, and Nayga 2013; Bello and Abdulai 2016), although some suggested that real values could be larger than the values constructed from hypothetical markets (Brownstone and Small 2005; Isacsson 2007; Hensher 2010). In some cases, researchers find no significant difference in valuations (i.e. estimated in terms of marginal WTPs) after comparing preferences between a hypothetical and a real choice experiment (Carlsson and Martinsson 2001; Lusk and Schroeder 2004).

Given the popularity of CEs in valuation research, it is important to find ways to mitigate hypothetical bias in CE studies. A number of approaches have been tested and used in the literature including cheap talk (Cummings and Taylor 1999; Silva et al. 2011; Mahieu, Riera, and Giergiczny 2012), oath (Carlsson et al. 2013; Jacquemet et al. 2013), consequentiality (Vossler, Doyon, and Rondeau 2012; Drichoutis et al. 2017), inferred valuation (Chang, Lusk, and Norwood 2009; Lusk and Norwood 2009; Norwood and Lusk 2011), and honesty priming (De-Magistris, Gracia, and Nayga 2013; Howard et al. 2017). Although methods such as inferred valuation, honesty priming, and consequentiality have been found effective in some empirical studies, concerns remain because adding complexity and additional validation could result in higher difficulty in comprehension and fatigue for the respondents (Bradley and Daly 1994; Cherchi and Hensher 2015). As DeShazo and Fermo (2002) shows, when a task exceeds the respondent's cognitive load, it could elicit inconsistent choices. Worse yet, people may also use simplifying strategies, such as skipping the alternatives or attributes, to ease choice task execution (Cherchi and Hensher 2015). Accordingly, visual aids such as pictures or images have been used in CEs in an attempt to reduce cognitive load and further enhance individuals' engagement in the choice tasks (Alfnes, Yue, and Jensen 2010; Van Wezemael et al. 2014; Cherchi and Hensher 2015).

Recent technology in virtual reality (VR) represents an important direction to potentially improve the validity of findings from CEs (Meißner and Decker 2010; Cherchi and Hensher 2015; Innocenti 2017). In particular, VR uses virtual environments to present digitally recreated real world activities via immersive (headsets) and non-immersive (2D computer screens) mediums (Parsons 2015). The intuition is simple: by providing controlled settings in which individuals decide and interact, VR environments "affect the dominant approach to laboratory experimentation in economics" (Innocenti 2017, p. 72). Moreover, immersive settings in VR are useful in investigating information processing and decision making adopted by individuals to reduce cognitive load and task complexity (Camerer, Loewenstein, and Prelec 2005; Sanchez-Vives and Slater 2005).

VR has been found effective in environmental economics studies aimed to improve consistency in choices by aiding the respondent's comprehension as well as enhancing their engagement and cognitive abilities (Bateman et al. 2009; Bonsall and Lythgoe 2009; Matthews, Scarpa, and Marsh 2017). For example, Bateman et al. (2009) found that compared to the standard presentation, preferences elicited in VR treatments are less variable and exhibit a significant reduction in asymmetry between willingness to pay and willingness to accept in the context of land use. Matthews, Scarpa, and Marsh (2017) indicated that VR reduced choice error, improved engagement, and resulted in different WTP from a static picture treatment. Despite the advantages of VR over the traditional delivery, however, no research has directly investigated whether VR can actually mitigate hypothetical bias in choice experiment. This paper fills this critical void in the literature.

The objective of this study is to assess whether the use of VR in CE can reduce or mitigate hypothetical bias. To examine this issue, we randomly assign subjects to different treatments that include hypothetical and non-hypothetical CEs. In the hypothetical CE treatments, subjects are presented with the choice sets using text, picture, or VR. In the non-hypothetical treatments, subjects are presented the choice sets in either text format or VR. Given the absence of real market for the new label at the time of the research, we approximate the real market with a binding choice in the non-hypothetical treatments. Using the context of food choices in a grocery store, our results suggest that the use of VR in CE does not completely eliminate, but can significantly reduce, hypothetical bias compared to computer based picture and plain-text delivery among participants who did not exhibit high VR discomfort.

General Design and Treatments' Description

Virtual Store

VR is a computer-mediated synthetic world that users can interactively experience through sensory stimuli (Fiore et al. 2009). Experiencing VR requires that the user has some means to navigate through the virtual world, which typically allow for interaction not just observation. The primary sense of vision is often dominated by having the user wear a head-mounted display, such as Oculus Rift^{TM} where the surrounding visual context is projected onto flat surfaces. The user's position and orientation are then tracked so that the visual experience is controlled by normal user movement and by gaze direction (Fiore et al. 2009). Navigation is achieved when the movements of the user are translated into movements within the virtual space.

As Innocenti (2017) noted, the key difference between VR experiments and standard laboratory tests is the higher level of immersion. Immersive settings are useful to investigate the processes of information processing and decision making adopted by individuals to reduce cognitive load and task complexity (Camerer, Loewenstein, and Prelec 2005; Sanchez-Vives and Slater 2005). The isolated experience in the headset allows people to focus solely on experimental stimuli (Lee et al. 2003; Fox, Bailenson, and Binney 2009). Such isolation allows focus on cognitive processes neglected in standard experiments (Innocenti 2017).

Our primary research question is whether an immersive virtual grocery store can enhance focus on the experimental stimuli and thereby mitigate hypothetical bias. To this end, a virtual grocery store was developed in order to customize an existing, highly immersive VR grocery store (Levy et al. 2015, 2019). The virtual store mimics a typical small-sized grocery store, providing similar grocery selection, shelf layout, entry, and exit. In the virtual grocery store, an illuminated pathway leads to the refrigeration section where yogurts were displayed. To implement our experiment in VR, we customized the presentation of the yogurt choice scenarios using the Unity VR software development platform (Unity.com, 2020). The labels were prepared based on the Food and Drug Administration's new guidelines (US Food and Drug Administration 2016) and scanned onto the virtual product. Figure 1 shows an example of the virtual store and choice environment. Oculus Rift^{TM} VR headsets were employed. This included wireless remote controls, at a unit cost of approximately $400. Each headset required a desktop computer to power the device and run the VR software. We then needed to retrofit our desktop machines with upgraded graphics cards (approximately $100 each) that were compatible with the Oculus^{TM} technology.

The VR experience was piloted to a small group of students first. Adjustments were then made based on feedback from the pilot. In particular, we added an audio voiceover that read the instructions. Respondents were guided on the use of the VR remote control and the complete range of motions that they could undertake within the virtual experience. They were free to take as long as they needed to arrive at the designated section and to pass by shelves filled with other groceries such as cereal and canned food. When arriving at the designated section, instruction appeared on how to navigate the controller in order to select a four-pack yogurt, view, zoom or rotate to see the nutrition labels. In order to approximate a realistic shopping experience, we asked respondents to enter the virtual store from a parking lot, push a shopping cart, navigate to the yogurt, complete the choice tasks, and then exit via the checkout counter. Moving avatars that represent other shoppers were present in the store but consistent for all VR participants. The experiment was recorded as soon as the participant entered the store and ended as he/she exited via the checkout counter, where he/she was greeted by an avatar clerk.

Spatial awareness acquired in a virtual environment has been shown to influence navigation and memory as participants who exhibit better viewing tend to have a higher sense of immersion (Wu et al. 2019). VR fully immerses the user in the virtual environment by blocking other visual inputs that might disturb the experience; however, this level of presence can lead to the undesirable simulator sickness (Tyrrell et al. 2018). Symptoms of simulator sickness resemble motion sickness to include eyestrain, headache, nausea, or even vomiting. Because simulator sickness could influence the cognitive transfer of knowledge, it is natural to assume that decision making would be influenced as well. To address the possibility of heterogeneous reactions to the simulated virtual store, we employed a questionnaire designed to measure the symptoms of simulator sickness by Kennedy et al. (1993), then modified by Tyrrell et al. (2018).

The simulator sickness questionnaire by Kennedy et al. (1993) is considered the current gold standard in simulator research. The questionnaire contained sixteen items on which subjects rate the degree of severity on a four-item scale (0 = none, 1 = slight, 2 = moderate, 3 = severe). The sixteen items can be summarized into three subscales: Nausea (general

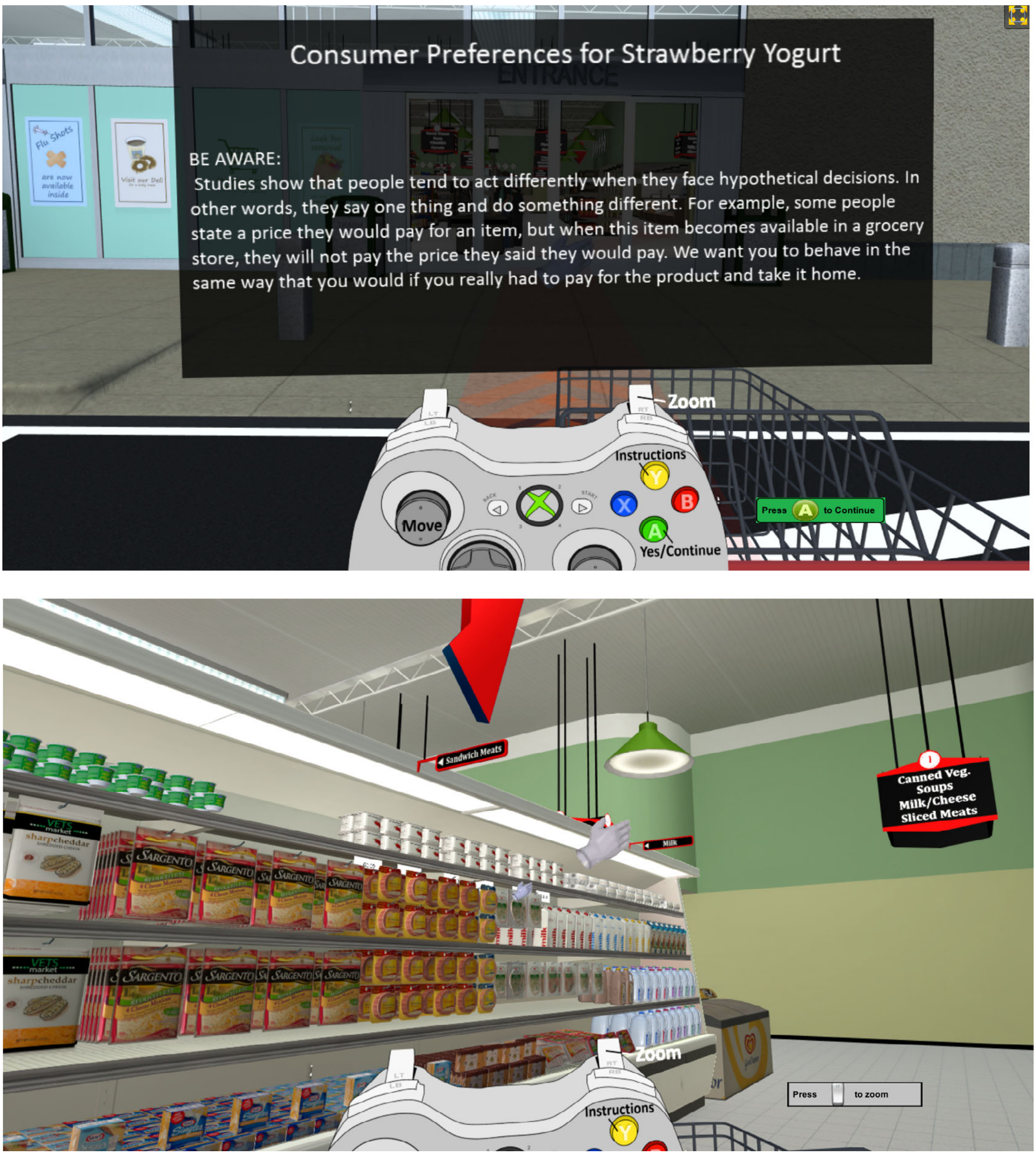

Figure 1. Choice experiment in an immersive virtual store [Color figure can be viewed at wileyonlinelibrary.com]

(Figure continues on next page.)

discomfort, increased salivation, sweating, nausea, difficulty concentrating, stomach awareness, burping), Oculomotor Stress (general discomfort, fatigue, headache, eyestrain, difficulty focusing, difficulty concentrating, blurred vision), and Disorientation (difficulty focusing, nausea, head fullness, blurred vision, dizzy, vertigo). We followed Tyrrell et al. (2018) to compose a weighted score, which reflects the total discomfort level as well as scores along the three subscales. This was calculated by summing the scores associated with each subscale and multiplying by an appropriate weighting factor: 9.54 for Nausea, 7.58 for Oculomotor, and 13.92 for Disorientation. Simulator sickness was measured before and after completing the choice tasks following Tyrrell et al. (2018). Because there is no standardized threshold to confirm "sickness", the overall scores before and after are often

Figure 1. (Continued)

compared to determine a change in comfort, which is indicated by a binary variable, with 1 indicating increased discomfort. To be consistent, this questionnaire was administered to participants in all treatments, not just the VR treatments. We make the entire VR survey available in Appendix 2 in Appendix S1.

Experimental Procedures

We follow Papoutsi et al. (2015) to employ a labeled design with three alternatives for each choice task: one strawberry yogurt labeled "Light," one strawberry yogurt labeled "Original," and a no-buy option labeled "Neither of these." The only other attribute is price. The design of our labeled CE is built upon the literature on the effect of nutritional labels on purchase decisions (Balcombea, Chalakb, and Fraser 2009; Petrovici et al. 2012). Four levels of the price attribute were specified to approximate the actual market prices for a four-cup (6 oz each) serving package of strawberry yogurt in the local stores: $1.65, $2.60, $3.55, and $4.50.

We employ a three-stage sequential Bayesian design to minimize the D-error following Bliemer, Rose, and Hess (2008) and Scarpa and Rose (2008). To obtain priors, we conducted one pilot with a sample of twenty-three students, implementing a fractional factorial orthogonal design. To determine the number of choice tasks, we followed ChoiceMetrics (2012) in order to satisfy the minimum number of degrees of freedom and attribute level balance.[1] To allow for more variation, we then opted for a design with eight choice tasks, with each choice task having three alternatives, that is, an original yogurt, a light yogurt, and a no-buy alternative. The eight choice tasks were divided in two blocks of four choice tasks each, meaning that each respondent faced four choice questions. This helps reduce fatigue in respondents, which has been considered an issue in CEs (Savage and Waldman 2008). The twenty-three responses from the pilot survey were used to estimate a multinomial logit (MNL) model in order to obtain mean values and standard errors of parameters, which were implemented as Bayesian priors assuming normal distributions. We used estimates from a MNL model because Bliemer and Rose (2010) demonstrated that designs based on MNL probabilities perform well for a random parameter logit model despite the difference in the asymptotic variance–covariance estimator. Interested readers can see similar examples in Bazzani et al. (2017). The advantage of such an approach is that the number of respondents necessary to reveal significant coefficients is reduced to a minimum. An example of the choice task is shown in figure 2.

Before the experiment, we provided all participants with detailed information regarding the experiment mechanism. In particular, they were informed that they would face four different choice tasks, each describing three alternatives. In the non-hypothetical treatments, respondents were informed that after responding to the choice tasks, one of the choice sets would be randomly chosen as the binding choice set. It was emphasized to the participants that each of the four choice tasks would have the same probability to be selected as binding. They would then be obligated to pay for the product they chose in the binding choice task if either of the yogurt options was chosen and to do so at the price indicated for the chosen product alternative. Actual yogurt labels with nutrition facts were created, then applied over the real light and original strawberry yogurt packages. Brand was not revealed. Upon completion, participants responded to a demographic questionnaire. A $10 participation fee was provided to each respondent at the end of the experiment.

We recruited students on the campus of a state university in the US. Respondents were randomly recruited among those who entered and exited the school cafeteria. They were informed of the opportunity to participate in a yogurt experiment rather than a VR experiment to avoid self-selection of tech-savvy people. When approaching the participants, interviewers asked them a set of screening questions: whether they were at least eighteen years old and whether they had purchased strawberry yogurt in the past six months. Eligible participants were then randomly assigned to one out of five between–subject treatments. In particular, we implemented three hypothetical CEs: plain text (HT),

[1]Namely, $S * (J - 1) >= K$, where S is the number of choice tasks, J is the number of the alternatives and K the number of parameters. Our experimental design consists of three alternatives in each choice set: original yogurt, light yogurt, and the no-buy option. Given our interest in preference heterogeneity, we utilized a random parameter logit model, specifying the two product alternatives (with the no-buy option as benchmark) and the price attribute to be random parameters. Accordingly, K is 6 because 6 parameters should be estimated: both the mean and standard deviation for each alternative. The minimum number of choice tasks necessary is then: $S >= 6/2 = 3$.

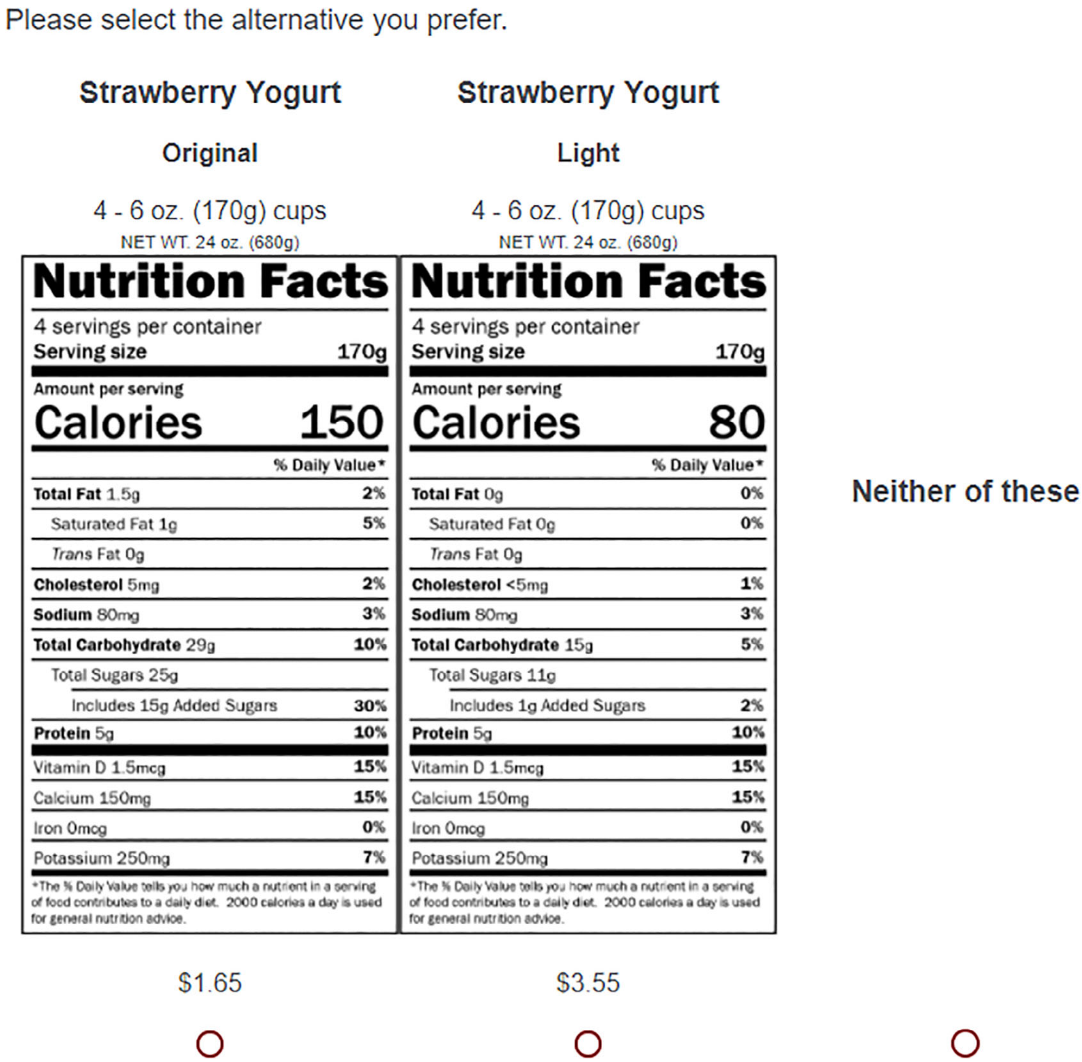

Figure 2. Sample choice task [Color figure can be viewed at wileyonlinelibrary.com]

pictures (HP) and, virtual grocery store (HVR). We conducted two non- hypothetical CEs: virtual grocery store (NVR) and plain text (NT). Approximately fifty respondents were randomly assigned to each treatment, resulting in a total sample of 256 subjects.

Hypotheses

To answer our research question, we build and test a series of hypotheses based on these treatments and total WTP estimates for original and light yogurt alternatives. First, we test whether the use of VR induces any "novel" effect. That is, people would behave similarly in non-hypothetical experiments, regardless of using VR or text. Hence, we expect WTP values from VR and text would be similar in the non-hypothetical CEs.

$$H0_1 : WTP^{NVR} - WTP^{NT} = 0, \text{ and}$$

$$H1_1 : WTP^{NVR} - WTP^{NT} > 0.$$

If we fail to reject $H0_1$, that suggests there is no difference between the two non-hypothetical CEs. Next, we test whether hypothetical bias indeed exists. Taking NT as the baseline, the next sets of hypotheses are defined as follows:

$$H0_2 : WTP^{HVR} - WTP^{NT} = 0, \text{ and}$$

$$H1_2 : WTP^{HVR} - WTP^{NT} > 0.$$

If $H0_2$ is rejected, we can confirm that hypothetical bias indeed exists in the hypothetical VR treatment.

$H0_3 : WTP^{HP} - WTP^{NT} = 0$, and

$H1_3 : WTP^{HP} - WTP^{NT} > 0$.

If $H0_3$ is rejected, we can confirm that hypothetical bias indeed exists in the hypothetical picture treatment.

$H0_4 : WTP^{HT} - WTP^{NT} = 0$, and

$H1_4 : WTP^{HT} - WTP^{NT} > 0$.

If $H0_4$ is rejected, we can confirm that hypothetical bias indeed exists in the hypothetical text treatment.

Next, we test whether VR produces smaller WTP values compared to other hypothetical treatments.

$H0_5 : WTP^{HVR} - WTP^{HP} = 0$,

$H1_5 : WTP^{HVR} - WTP^{HP} < 0$.

If $H0_5$ is rejected, we can confirm that the WTP values produced in HVR are smaller than those in HP.

$H0_6 : WTP^{HVR} - WTP^{HT} = 0$, and

$H1_6 : WTP^{HVR} - WTP^{HT} > 0$.

If $H0_6$ is rejected, we can confirm that the WTP values produced in HVR are smaller than those in HT.

As we described in *Experimental Procedures*, we conducted a between-subject experiment where respondents are randomly assigned to the different treatments. Therefore, even though we can test the difference in WTPs between two treatments, we cannot test the difference in WTP between two treatments against that of another two treatments because no participant has conducted more than one treatment. Given this, we cannot directly examine the reduction of WTP in HVR against that in HP or HT with hypothesis testing, i.e., $H_0 : WTP^{HVR} - WTP^{NT} = WTP^{HP} - WTP^{NT}$. We do so in the regression analysis specified next.

Model

Our model follows the one proposed by Fiebig et al. (2010). We start with the standard specification where the utility of person n from choosing alternative j in choice set t is:

$$(1) \quad U_{njt} = \beta x_{njt} + \varepsilon_{njt},$$

where x_{njt} is a vector of the observed attributes of alternative j; β is a vector of mean attribute utility weights and homogeneous across consumers, and ε_{njt} is well-behaved (e.g. IIA) idiosyncratic error term. It is apparent that such a naïve model can be easily violated. Therefore, we add to the standard specification by allowing a vector of person n-specific deviations from the mean β; that is, $U_{njt} = (\beta + \eta_n) x_{njt} + \varepsilon_{njt}$, where η_n is a vector of person specific taste difference from the mean based on observed attributes. To further address consumer heterogeneity in tastes for unobserved product attribute, we add a scale parameter, σ, to the idiosyncratic error term, that is, $U_{njt} = (\beta + \eta_n)x_{njt} + \varepsilon_{njt}/\sigma$. This specification is summarized in Fiebig et al. (2010) as a generalized multinomial mixed logit model (GMNL):

$$(2) \quad U_{njt} = [\sigma\beta + \gamma\eta_n + (1-\gamma)\sigma\eta_n]x_{njt} + \varepsilon_{njt},$$

where γ is bounded by [0, 1] and indicates how the variance of residual taste heterogeneity varies with the scale σ. To illustrate, when $\gamma = 1$, $\beta_n = \sigma\beta + \eta_n$, the variance is independent of the scale parameter related to the idiosyncratic term. When $\gamma = 0$, $\beta_n = \sigma(\beta + \eta_n)$, taste heterogeneity is proportional to the scale parameter. Assuming the scale parameter is heterogeneous (person specific) in the population, σ_n leads to the calculation of individual WTP values. We made two assumptions about the individual scale parameter. First, we let it follow a log-normal distribution with mean 1 and standard deviation τ, that is, $\sigma_n \sim \text{LN}(1, \tau^2)$. This assumes that consumer heterogeneity is random. Second, we let σ_n follow a function of known predictors, where we assume consumer heterogeneity is determined by reported consumer characteristics. In this case, consumer heterogeneity is assumed to be influenced by self-reported health status, nutrition label use, simulator sickness, and nutrition attitudes. The focus here is not to estimate these factors separately but to take these as inputs to better predict the scale parameter, σ_n.[2] The first assumption leads to Model 1 and the second assumption leads Model 2. Following Fiebig

[2]For a similar application with more discussion on the scale parameter, please refer to Fang et al. (2019).

et al. (2010), the probability of person n choosing alternative j under choice set t is:

$$(3) \quad Pr(j|x_{nt}|)$$

$$= \frac{1}{D}\sum_{d=1}^{D}\frac{\exp\left(\sigma^d\beta+\gamma\eta^d+(1-\gamma)\sigma^d\eta^d\right)x_{njt}}{\sum_{k=1}^{J}\exp(\sigma^d\beta+\gamma\eta^d+(1-\gamma)\sigma^d\eta^d)x_{nkt}},$$

where d is the number of random draws. The likelihood function is:

$$(4) \quad L(\gamma,\tau|x_{nt}) = \frac{1}{D}\sum_{1}^{D}\prod_{t}\prod_{j}P(j(x_{nt}))^{y_{njt}}$$

$$= \frac{1}{D}\sum_{d=1}^{D}\prod_{t}\prod_{j}\left(\frac{\exp\left(\sigma^d\beta+\gamma\eta^d+(1-\gamma)\sigma^d\eta^d\right)\ x_{njt}}{\sum_{k=1}^{J}\exp(\sigma^d\beta+\gamma\eta^d+(1-\gamma)\sigma^d\eta^d)\ x_{nkt}}\right)^{y_{njt}}.$$

where y_{njt} is the decision (1 or 0). Our models were estimated using the *gmnl* Package (Sarrias and Daziano 2017). Results presented in this manuscript were produced with Rstudio version 1.2.5019. Replication codes and de-identified datasets can be found in Appendix 3 in Appendix S1.

We allowed for random effects in normal distribution for *light* and *original*, and a log-normal distribution for *price*. We also allowed our alternative specific constant (*ASC*) for each attribute to be random following a normal distribution and scaled by σ_{nt}.[3] In our empirical specification, components of the utility function include price, the different treatments, and an alternative-specific option representing either original or the light option:

$$(5) \quad U_{njt} = [\sigma_n\beta+\gamma\eta_n+(1-\gamma)\sigma_n\eta_n](ASC_{original}+ASC_{light}+Price+ASC_{original}\times NT+ ASC_{light}\times NT+ASC_{original}\times HT+ASC_{light}\times HT+ASC_{original}\times HP+ ASC_{light}\times HP+ASC_{original}\times HVR+ASC_{light}\times HVR+\varepsilon_{njt}$$

Coefficient for *Price* was fixed at −1 and we used the scale parameter σ_{nt} to adjust for individual heterogeneity. The WTP for each alternative was estimated directly as:

$$(6) \quad WTP_{njt} = \sigma_n\beta+\gamma\eta_n+(1-\gamma)\sigma_n\eta_n,$$

rather than a ratio of the mean parameter estimated for the non-monetary attributes to the mean price parameter. In other words, our WTP values are the regression coefficients. Greene and Hensher (2010) recommended fixing γ at 0 to estimate the model in the "WTP space." Although doing so sacrifices the flexibility of the model, it may reduce the incidence of exceedingly large or small WTP (Greene and Hensher 2010). Hence, we also estimated the models with γ equal to 0 as a robustness check in Appendix 1 in Appendix S1.

Results

Table 1 exhibits the overall descriptive statistics of our sample and the balance among treatments with respect to the socio-demographic variables. Two hundred and fifty six participants were recruited from a land-grant university, with approximately fifty participants for each treatment.[4] Each participant completed four choice tasks. Using a student sample is not uncommon in CE because student behavior in a non-hypothetical valuation study has been found to be similar to that of non-students (Depositario et al. 2009). Even though a student sample may exhibit certain skewed behaviors, it should not bias our conclusion given our interest in the differences in results across treatments where respondents

[3]Fiebig et al. (2010) argued that allowing *ASC*s to be scaled could lead to overly large estimates. However, the alternative is to assume that consumers are homogeneous toward the "no-buy" option, which we deem to be unrealistic. Hence, we allow the residual taste to be heterogeneous for the "no-buy" option because some consumers may opt out due to undesirable product attributes, whereas others may opt out because they are unwilling to purchase a strawberry yogurt.

[4]This experiment is approved by the University of Arkansas IRB protocol# 17-04-623.

Table 1. Descriptive Statistics and Balance Tests

Treatments	NT	NVR	HVR	HT	HP	χ^2, F	p-Value
Observations	49	51	51	51	54.00		
Age						12.194	0.143
18–19	24.50%	35.30%	45.10%	41.20%	40.70%		
20–21	40.80%	45.10%	21.60%	31.40%	25.90%		
22+	34.70%	19.60%	33.30%	27.50%	33.30%		
Gender						0.780	0.941
Male	53.10%	58.80%	51.00%	52.90%	51.90%		
Female	46.90%	41.20%	49.00%	47.10%	48.10%		
Household income						13.643	0.324
0–29,999	49.00%	35.30%	43.10%	29.40%	37.00%		
30,000–59,999	12.20%	23.50%	15.70%	29.40%	16.70%		
60,000–89,999	10.20%	23.50%	15.70%	11.80%	14.80%		
90,000+	28.60%	17.60%	25.50%	29.40%	31.50%		
Class						15.612	0.480
Freshman	14.30%	25.50%	13.70%	29.40%	24.10%		
Sophomore	30.60%	21.60%	31.40%	13.70%	25.90%		
Junior	22.40%	29.40%	21.60%	23.50%	16.70%		
Senior	22.40%	11.80%	17.60%	19.60%	13.00%		
Other	10.20%	11.80%	15.70%	13.70%	20.40%		
Reads nutrition facts label						5.236	0.264
No	34.70%	21.60%	35.30%	31.40%	20.40%		
Yes	65.30%	78.40%	64.70%	68.60%	79.60%		
Nutrition importance	3.612 (1.057)	3.706 (1.171)	3.392 (1.150)	3.608 (1.150)	3.778 (1.110)	0.859	0.489
Health score	5.374 (1.011)	5.034 (1.276)	5.282 (0.959)	5.264 (1.169)	5.168 (1.221)	0.654	0.625
Sickness score (post)	122.900 (167.530)	172.834 (201.446)	210.769 (290.210)	133.229 (346.985)	152.195 (208.652)	0.972	0.423
Sickness score (binary)	18.40%	39.20%	58.80%	17.60%	16.70%	33.623	<0.001

Table 2. Summary of Simulator Sickness

SS Questionnaire		Subscales	NT				HVR				HT				HP			
			before mean	Std. dev.	After mean	Std. dev.	before mean	Std. dev.	After mean	Std. dev.	before mean	Std. dev.	After mean	Std. dev.	before mean	Std. dev.	After mean	Std. dev.
SSQ1	General discomfort	N, O	0.143	0.350	0.143	0.350	0.118	0.378	0.294	0.604	0.176	0.513	0.196	0.525	0.167	0.500	0.148	0.488
SSQ2	Fatigue	O	0.347	0.556	0.286	0.495	0.333	0.583	0.235	0.469	0.431	0.748	0.294	0.695	0.519	0.811	0.426	0.736
SSQ3	Headache	O	0.122	0.385	0.122	0.385	0.118	0.378	0.196	0.486	0.294	0.636	0.275	0.660	0.278	0.559	0.259	0.551
SSQ4	Eye strain	O	0.163	0.370	0.143	0.350	0.098	0.298	0.294	0.535	0.216	0.536	0.157	0.364	0.370	0.618	0.278	0.525
SSQ5	Difficulty focusing	O, D	0.510	0.732	0.449	0.785	0.176	0.430	0.196	0.561	0.196	0.486	0.118	0.471	0.426	0.736	0.370	0.702
SSQ 6	Salivation increasing	N	0.143	0.350	0.224	0.545	0.216	0.536	0.235	0.546	0.216	0.498	0.294	0.604	0.167	0.420	0.222	0.458
SSQ7	Sweating	N	0.327	0.652	0.224	0.545	0.059	0.235	0.137	0.344	0.137	0.344	0.118	0.471	0.315	0.604	0.259	0.551
SSQ8	Nausea	N, D	0.020	0.142	0.000	0.000	0.176	0.585	0.431	0.748	0.235	0.614	0.216	0.605	0.111	0.416	0.093	0.398
SSQ 9	Difficulty concentrating	N, O	0.469	0.786	0.408	0.727	0.137	0.397	0.216	0.536	0.255	0.622	0.176	0.513	0.333	0.667	0.352	0.699
SSQ10	Fullness of the head	D	0.082	0.274	0.061	0.240	0.078	0.389	0.235	0.469	0.235	0.645	0.157	0.538	0.130	0.433	0.130	0.387
SSQ11	Blurred vision	O, D	0.020	0.142	0.000	0.000	0.078	0.269	0.157	0.459	0.137	0.525	0.098	0.454	0.130	0.387	0.130	0.387
SSQ12	Dizziness with eyes open	D	0.020	0.142	0.020	0.142	0.039	0.194	0.471	0.751	0.098	0.358	0.078	0.436	0.000	0.000	0.019	0.135
SSQ13	Dizziness with eyes closed	D	0.000	0.000	0.020	0.142	0.039	0.194	0.314	0.577	0.157	0.459	0.118	0.471	0.019	0.135	0.037	0.189
SSQ14	Vertigo	D	0.000	0.000	0.000	0.000	0.020	0.139	0.157	0.500	0.059	0.308	0.078	0.436	0.074	0.325	0.074	0.325
SSQ15	Stomach awareness	N	0.429	0.639	0.347	0.591	0.373	0.816	0.451	0.776	0.333	0.733	0.373	0.685	0.278	0.559	0.333	0.578
SSQ16	Burping	N	0.082	0.274	0.061	0.240	0.157	0.500	0.020	0.139	0.039	0.278	0.059	0.416	0.037	0.189	0.056	0.229

Note: N indicates Nausea, O indicates Oculomotor stress, and D indicates Disorientation.

Table 3. Main Results

	Model 1		Model 2	
Parameter	estimate	Std. err.	estimate	Std. err.
Original				
ASC	2.743***	0.142	2.589***	0.133
sd. (ASC)	2.818***	0.639	2.891***	0.576
ASC × HVR	0.635***	0.242	0.762***	0.269
ASC × HT	1.058***	0.328	1.017***	0.306
ASC × HP	0.878***	0.274	0.821***	0.213
Light				
ASC	2.999***	0.167	3.013***	0.154
sd. (ASC)	2.498***	0.653	2.524***	0.598
ASC × HVR	0.862***	0.25	0.756***	0.273
ASC × HT	0.977***	0.285	0.806***	0.264
ASC × HP	1.562***	0.3	1.383***	0.264
Consumer heterogeneity				
γ	1.684***	0.188	1.493***	0.123
τ	1.143***	0.336	1.439***	0.219
Intercept	1.294***	0.442	2.494***	0.563
Health score			−0.301***	0.128
Nutrition importance			0.350***	0.121
Label use			−0.543**	0.269
Simulator sickness (binary)			−0.839***	0.174
N	820		820	
LL	−519.74		−523.56	
AIC	1065.48		1081.12	

Note: ***, **, *Indicates significance levels at 0.01, 0.05, and 0.10. Model 1 allows the scale parameter to be drawn from a log-normal distribution. Model 2 estimates the scale parameter based on Health Score, Nutrition Importance, Label Use, and Simulator Sickness. The baseline group is non-hypothetical text (NT). Coefficients are WTP.

are randomly assigned. As illustrated by the χ^2 tests, we achieved balance across treatments in reported characteristics in age, gender, grade, and income (p-value > 0.05). Balance was also achieved regarding label use, self-reported health status as well as nutrition conscientiousness. Overall, the majority of participants were undergraduate students between the age of 18 and 22. A slightly larger proportion (58.80%) of the participants were male. A majority (71.32%) of participants reported reading the label when they shop for yogurts. The self-reported nutrition importance is 3.62 out of 5, and the self- reported health status is 5.28 out of 7, for all participants, indicating that our sample has a higher consideration for nutrition and health.

We also examined responses to the simulator sickness questionnaire. When considering only the post-experiment simulator sickness score (SSC), or the continuous measure, there is no difference across treatments. The majority of those in the non-VR treatments do not experience any discomfort. In the two VR treatments, however, approximately twenty in non-hypothetical and thirty in hypothetical CE reported an increased SSC post-experiment. Therefore, we created a binary indicator for those who experienced increased overall SSC post experiment. This measure is later used to exclude participants in a robustness check. As previously mentioned, the questionnaire includes sixteen items in three subscales. Table 2 shows the average score for each item in the questionnaire before and after the experiment.

Table 3 presents the main results. Model 1 allows the scale to be random following a log-normal distribution whereas Model 2 accounts for consumer heterogeneity determined by known simulator sickness, label use, self-reported health status, and nutrition conscientiousness. For Model 2 in particular, we let $\sigma_n = exp(\sigma + \delta_1 SimSickness + \delta_2 HealthScore + \delta_3 LabelUse + \delta_4 NutritionImp + \tau\varepsilon_0)$, where *SimSickness* indicates the discomfort induced by the VR treatment, measured by a binary indicator of greater weighted overall score post experiment. To explore the possibility that people who are health oriented may exhibit different preference, we followed Pieniak, Verbeke, and Scholderer (2010) and created an index using agreement responses (seven-point Likert scale) to eleven statements about the importance of health (e.g., health is

Table 4. Robustness Check on Simulator Sickness

	Model 3		Model 4	
	estimate	Std. error	estimate	Std. error
Original				
ASC	2.755***	0.158	2.681***	0.141
Sd. (ASC)	2.867***	0.577	3.359***	0.470
ASC × HVR	0.659*	0.280	0.634	0.373
ASC × HVR × SSC	0.002	0.002		
ASC × HT	0.934**	0.362	1.054**	0.358
ASC × HT × SSC	0.001	0.002		
ASC × HP	0.638**	0.247	0.886**	0.309
ASC × HP × SSC	−0.001	0.001		
Light				
ASC	3.117***	0.174	3.042***	0.181
sd. ASC	2.617***	0.624	2.380***	0.466
ASC × HVR	0.777***	0.284	0.844*	0.372
ASC × HVR × SSC	0.001	0.002		
ASC × HT	0.657*	0.287	1.012**	0.334
ASC × HT × SSC	−0.002	0.001		
ASC × HP	1.283***	0.281	1.420***	0.322
ASC × HP × SSC	0.000	0.001		
Consumer heterogeneity				
γ	1.560***	0.145	1.021***	0.125
τ	1.151***	0.197	0.955***	0.184
Intercept	2.763***	0.507	2.336***	0.579
Health score	−0.536***	0.075	−0.229***	0.086
Nutrition importance	0.379***	0.108	−0.115	0.095
Label use	0.033	0.235	0.433	0.225
N. obs	820		700	
LL	−508.66		−453.28	
AIC	1085.02		928.79	

Note: ***, **, *Indicates significance levels at 0.01, 0.05, and 0.10. Model 3a estimates the scale parameter based on Health Score, Nutrition Importance, and Label Use. Model 3a includes interactions of the post-experiment SSC with each treatment. Model 3b excludes people who are simulator sick and estimates the scale parameter based on Health Score, Nutrition Importance, and Label Use. The baseline group is non-hypothetical text (NT). Coefficients are WTP.

very important to me). The eleven values were then averaged to create *HealthScore*. Furthermore, unfamiliarity with nutrition label could diminish the effects induced by the new label, as those who are unfamiliar may display a lower degree of understanding toward nutrition content and be less responsive to new details than those who frequently use labels. To explore this, each participant responded either "never," "infrequently," "frequently," or "very frequently" to the following question as part of the survey questionnaire: "How often do you read nutrition information on food labels when purchasing food products?" Categorical values were then created for the variable *LabelUse*. Finally, we explored whether individuals who self-report to be more nutrition conscious exhibit different preferences compared to individuals who are less nutrition-conscious. Using a five-point Likert scale, participants responded to the following question: "How important to you is nutritional value when making choices concerning yogurt products?" The responses were then coded for the *NutritionImp* variable.

Because all factors in the scale parameter are statistically significant, we consider Model 2 as the preferred model. The baseline treatment is NT, therefore its WTP is represented by the *ASC* intercept. The WTP values for other treatments are calculated by adding the baseline coefficient and the coefficients of the interactions. Not surprisingly, the WTP in NT produced the smallest values: $2.589 for original yogurt and $3.013 for light yogurt. This is consistent with previous literature that consumers tend to overstate their values in hypothetical experiments (List and Gallet 2001; Murphy et al. 2005; Chowdhury et al. 2011; De-Magistris, Gracia, and Nayga 2013). Interaction terms indicate the difference between NT and other hypothetical treatments. For original yogurt, participants overstated their WTP by $0.762 in HVR, $1.017 in HT, and

Table 5. Hypotheses Testing on Hypothetical Bias

		Conditional WTP		Unconditional WTP		
		t-Statistic	p-Value	t-Statistic	p-Value	Conclusion
H1	Light	0.972	0.167	0.263	0.397	Accept
H1	Original	–1.639	0.948	–0.578	0.718	Accept
H2	Light	9.506	<0.001	5.237	<0.001	Reject
H2	Original	6.781	<0.001	4.943	<0.001	Reject
H3	Light	15.774	<0.001	10.079	<0.001	Reject
H3	Original	6.989	<0.001	5.232	<0.001	Reject
H4	Light	15.738	<0.001	10.546	<0.001	Reject
H4	Original	10.479	<0.001	8.083	<0.001	Reject
H5	Light	9.776	<0.001	9.675	<0.001	Reject
H5	Original	8.289	<0.001	8.640	0.003	Reject
H6	Light	12.450	<0.001	12.296	<0.001	Reject
H6	Original	10.965	<0.001	12.271	<0.001	Reject

Note: The baseline group is NT. Conditional t test uses WTP values from Model 2, and unconditional t test uses WTP values that were bootstrapped from the normal distribution based on estimates from Model 2.

$0.821 in HP. For light yogurt, participants overstated their WTP by $0.756 in HVR, $0.806 in HT, and $1.383 in HP. These results indicate that VR produced the smallest bias.

Table 4 presents the robustness check on simulator sickness. The binary indicator of simulator sickness is significant in the scale parameter, indicating that simulator sickness may have influenced individual preference. We further conduct a robustness check using a continuous post-experiment SSC instead of a binary sickness indicator. In Model 3, we included interactions of the continuous post-experiment SSC with each treatment. The results are consistent in that HVR produced the smallest WTP in all hypothetical treatments. All the interaction terms are nonsignificant. We then take a more radical approach and excluded thirty participants in VR who exhibited higher discomfort level after the experiment.[5] These results are presented in Model 4. Not only did HVR produce the smallest values, there was also no statistically significant difference between HVR and NT for original yogurt. Therefore, we conclude that VR performed well in reducing hypothetical bias when controlling for simulator sickness.

Of the four factors examined for consumer heterogeneity, three were found to be significant. Specifically, it is not surprising that health attitude, self-reported health status, and frequency of label use influence individual preferences for yogurt purchase. This is consistent with previous studies in the food choice literature (Nayga, Lipinski, and Savur 1998; Drichoutis, Lazaridis, and Nayga 2005; Hess, Visschers, and Siegrist 2012; Hammond et al. 2015). Note that we cannot meaningfully interpret the coefficients in the model because they are used to estimate the scale parameter, not the purchase choice. As previously mentioned, Greene and Hensher (2010) recommended fixing γ at 0 to estimate the model in the "WTP space." Hence, we estimated the models with γ equal to 0, and the results are presented in Appendix 1in Appendix S1. The results are largely consistent.

Discussion of Hypothetical Bias

We estimated both the conditional and unconditional individual WTPs. The latter were bootstrapped from the normal distribution based on estimates from Model 2. We then applied t-tests to test for significant differences in terms of conditional and unconditional WTP values across the treatments. The results are shown in table 5.

Failure to reject the null hypothesis of H1 leads to the conclusion that there is no significant difference between the two non-hypothetical CEs. In other words, VR does not seem to induce a novel effect compared to text treatment. Once we confirm that there is no significant difference between the two non-hypothetical CEs, we choose to use NT as the baseline. Next, we test

[5]Note this exclusion criterion is rigorous because we removed people who experienced only slight discomfort as well. We also sacrificed some power in this case because we are removing 60% of the VR sample. However, since each participant answered four choice sets, our number of observations is still 80 in the VR sample. The fact that we are able to detect significant and consistent effects with a smaller sample shows that our results are robust regardless of simulator sickness.

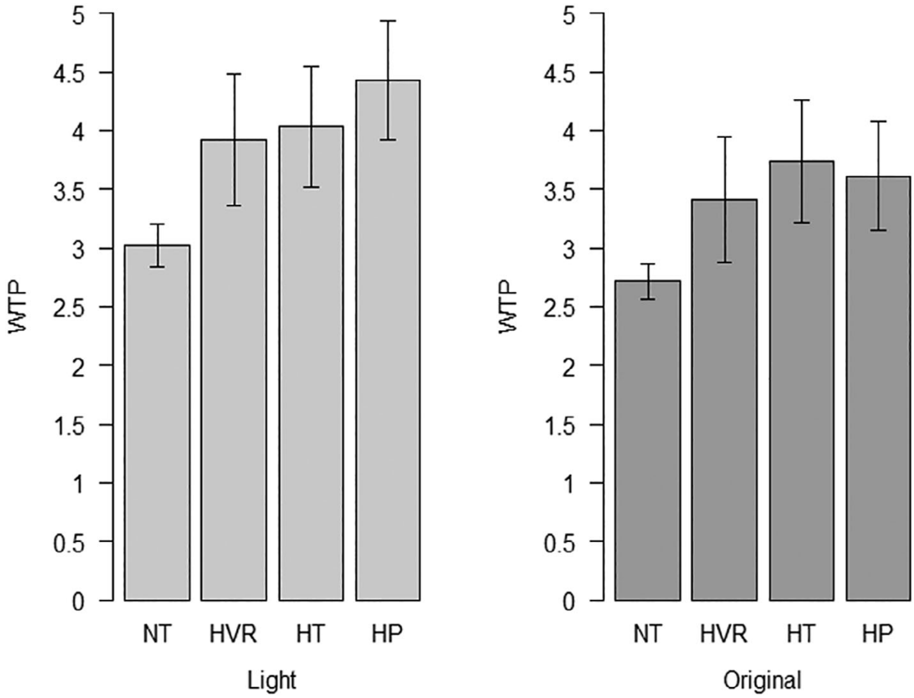

Figure 3. WTP for all treatments

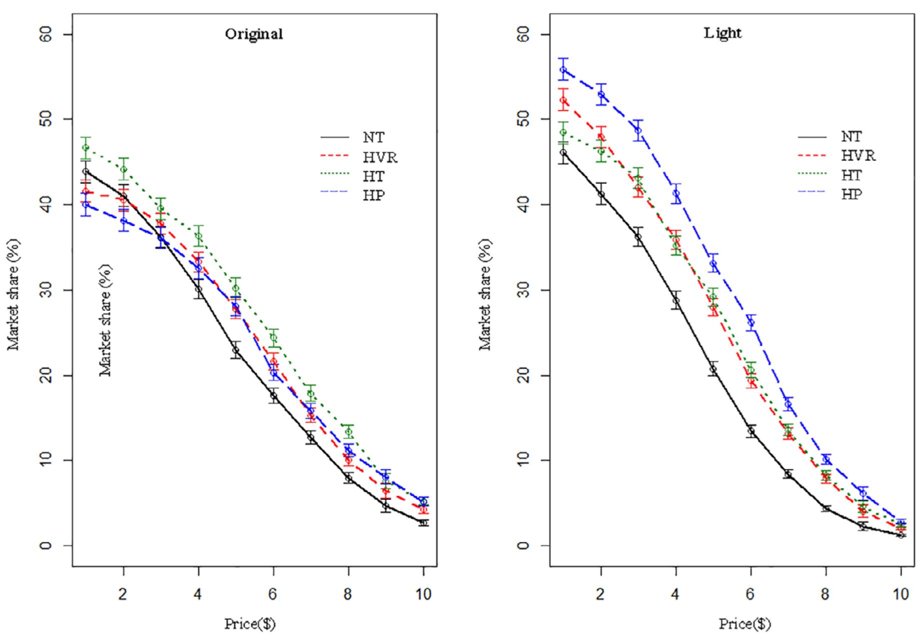

Figure 4. Simulated market shares [Color figure can be viewed at wileyonlinelibrary.com]

whether the WTP values are significantly larger in each hypothetical treatment compared to NT, using one-sided tests through H2–H4. We reject the null hypotheses in favor of the alternatives, concluding the existence of hypothetical bias in VR, picture, and text treatments.

Last, we examine the differences in WTP values between HVR and HP and between

HVR and HT. We show that HVR produced a smaller WTP compared to HP and HT by rejecting the null hypotheses of H5 and H6. As discussed earlier, we cannot directly test the magnitude of reduction using a two-sample t test due to the between-subject experiment design. We turn to the regression results in Model 2 for some additional insights. Results show that although VR does not completely eliminate hypothetical bias, it does reduce it. To visualize this finding, WTP values are shown in figure 3 to illustrate the results from Model 2. The 95% confidence intervals are also exhibited on the graph. Clearly, NT presents the lowest WTP, and among the hypothetical treatments, VR produces the lowest WTP.

To illustrate how sensitive our conclusion is to a price change, we plot the predicted probabilities of purchase based on Model 2 estimates. Figure 4 shows the simulated market shares. The NT market share is clearly lower than all hypothetical market shares at all price levels. Importantly, we also observe that hypothetical VR is closest to NT at every price level. Overall, hypothetical VR has the lowest market share values among all the hypothetical treatments.

Conclusion

Choice experiments are widely used by economists and other social scientists in valuing private and public goods. Like other stated preference methods, however, it has been shown to produce hypothetical bias in WTP estimates. Hence, finding ways to mitigate this hypothetical bias in CE studies is important. This study tests the use of VR in reducing hypothetical bias in CE studies. To accomplish this, we randomly assigned participants to several treatments that include three hypothetical CEs that differed on the way the choice sets are presented (i.e. text, picture, and VR) and two non-hypothetical CEs (i.e. text and VR). Using the non-hypothetical text treatment as the benchmark, our results suggest that hypothetical bias existed in all the hypothetical treatments but that VR produces the smallest WTP values. This is especially true when we excluded participants who experienced simulator sickness. Hence, our results suggest that VR may mitigate hypothetical bias to some degree.

An implication of this important finding is that lack of realism could be one of the reasons for the occurrence of hypothetical bias in hypothetical choice experiments, and so the addition of some form of realism in the form of immersion to a virtual world using VR technology could add some realism to the choice environment. A word of caution is, however, warranted. The fact that discomfort in a virtual environment could alter behavioral pattern is worth noting because the advantage of VR is that it can block out other visual cues that could cause cognitive overload. However, such "blocking" could also lead to simulator sickness. Our results suggest that this should be taken into account when utilizing and analyzing data using VR.

Given the evidence pointing to the existence of hypothetical bias in stated preference studies, non-hypothetical versions of CE have emerged in the literature. It is, however, not always possible to conduct CE studies non-hypothetically in real markets (Chavez et al. 2020). Hence, the fact that we found VR, after considering simulator sickness, may mitigate hypothetical bias is important. Future research should, however, replicate our study in other contexts to test the robustness of our results. It would also be interesting in the future to compare results from the use of VR vis-a-vis other ex-ante techniques that have been used in CE studies to potentially reduce hypothetical bias (e.g. cheap talk, honesty priming, consequentiality, inferred valuation, oath). Our hope is that this paper will encourage and motivate researchers using stated preference methods to try VR to potentially reduce hypothetical bias.

Acknowledgement

The authors would like to thank Mohan Zalake, Shivashankar Halan, PhD, and Michael Borish, PhD, who served as graduate assistants for Dr. Lok, for their help in developing the virtual store.

Supplementary Material

Supplementary material are available at *American Journal of Agricultural Economics* online.

References

Alfnes, Frode, Chengyan Yue, and Helen H Jensen. 2010. Cognitive Dissonance as a

Means of Reducing Hypothetical Bias. *European Review of Agricultural Economics* 37: 147–63.

Balcombea, Kelvin, Ali Chalakb, and Iain Fraser. 2009. Model Selection for the Mixed Logit with Bayesian Estimation. *Journal of Environmental Economics and Management* 57(2): 226–37.

Bateman, Ian J, Brett H Day, Andrew P Jones, and Simon Jude. 2009. Reducing Gain-Loss Asymmetry: A Virtual Reality Choice Experiment Valuing Land Use Change. *Journal of Environmental Economics and Management* 58(1): 106–18.

Bazzani, Claudia, Vincenzina Caputo, Rodolfo M Nayga JR, and Maurizio Canavari. 2017. Testing Commitment Cost Theory in Choice Experiments. *Economic Inquiry* 55: 383–96.

Bello, Muhammad, and Awudu Abdulai. 2016. Impact of Ex-ante Hypothetical Bias Mitigation Methods on Attribute Non-attendance in Choice Experiments. *American Journal of Agricultural Economics* 98: 1486–506.

Bliemer, Michiel C, John M Rose, and Stephane Hess. 2008. Approximation of Bayesian Efficiency in Experimental Choice Designs. *Journal of Choice Modelling* 1: 98–126.

Bliemer, Michiel CJ, and John M, Rose. 2010. Construction of experimental designs for mixed logit models allowing for correlation across choice observations. *Transportation Research Part B: Methodological* 44(6): 720–734.

Bonsall, Peter, and Bill Lythgoe. 2009. Factors Affecting the Amount of Effort Expended in Responding to Questions in Behavioural Choice Experiments. *Journal of Choice Modelling* 2: 216–36.

Bradley, Mark, and Andrew Daly. 1994. Use of the Logit Scaling Approach to Test for Rank-Order and Fatigue Effects in Stated Preference Data. *Transportation* 21: 167–84.

Brownstone, David, and Kenneth A Small. 2005. Valuing Time and Reliability: Assessing The Evidence from Road Pricing Demonstrations. *Transportation Research Part A: Policy and Practice* 39: 279–93.

Camerer, Colin, George Loewenstein, and Drazen Prelec. 2005. Neuroeconomics: How Neuroscience Can Inform Economics. *Journal of Economic Literature* 43: 9–64.

Carlsson, Fredrik, Mitesh Kataria, Alan Krupnick, Elina Lampi, Åsa Löfgren, Ping Qin, and Thomas Sterner. 2013. The Truth, the Whole Truth, and Nothing but the Truth: A Multiple Country Test of an Oath Script. *Journal of Economic Behavior & Organization* 89: 105–21.

Carlsson, Fredrik, and Peter Martinsson. 2001. Do Hypothetical and Actual Marginal Willingness to Pay Differ in Choice Experiments?: Application to the Valuation of the Environment. *Journal of Environmental Economics and Management* 41 (2): 179–92.

Chang, Jae B, Jayson L Lusk, and F Bailey Norwood. 2009. How Closely Do Hypothetical Surveys and Laboratory Experiments Predict Field Behavior? *American Journal of Agricultural Economics* 91: 518–34.

Chavez, Daniel E, Marco A Palma, Rodolfo M Nayga, and James W Mjelde. 2020. Product Availability in Discrete Choice Experiments with Private Goods. *Journal of Choice Modelling*: 100225. https://doi.org/10.1016/j.jocm.2020.100225.

Cherchi, Elisabetta, and David A Hensher. 2015. Workshop Synthesis: Stated Preference Surveys and Experimental Design, an Audit of the Journey so Far and Future Research Perspectives. *Transportation Research Procedia* 11: 154–64.

ChoiceMetrics. 2012. Ngene 1.1. 1 User Manual & Reference Guide. *ChoiceMetrics* 19: 20.

Chowdhury, Shyamal, J Meenakshi, Keith I Tomlins, and Constance Owori. 2011. Are Consumers in Developing Countries Willing to Pay More for Micronutrient-Dense Biofortified Foods? Evidence from a Field Experiment in Uganda. *American Journal of Agricultural Economics* 93: 83–97.

Cummings, Ronald G, and Laura O Taylor. 1999. Unbiased Value Estimates for Environmental Goods: a Cheap Talk Design for The Contingent Valuation Method. *American Economic Review* 89: 649–65.

De-Magistris, Tiziana, Azucena Gracia, and Rodolfo M Nayga. 2013. On the Use of Honesty Priming Tasks to Mitigate Hypothetical Bias in Choice Experiments. *American Journal of Agricultural Economics* 95: 1136–54.

Depositario, Dinah Pura T, Rodolfo M Nayga, Ximing Wu, and Tiffany P Laude. 2009.

Should Students Be Used as Subjects in Experimental Auctions? *Economics Letters* 102: 122–4.

DeShazo, JR, and German Fermo. 2002. Designing Choice Sets for Stated Preference Methods: The Effects of Complexity on Choice Consistency. *Journal of Environmental Economics and Management* 44(1): 123–43.

Drichoutis, Andreas C, Panagiotis Lazaridis, and Rodolfo M Nayga. 2005. Nutrition Knowledge and Consumer Use of Nutritional Food Labels. *European Review of Agricultural Economics* 32: 93–118.

Drichoutis, Andreas C, Achilleas Vassilopoulos, Jayson L Lusk, and Rodolfo M Nayga. 2017. Consumer Preferences for Fair Labour Certification. *European Review of Agricultural Economics* 44: 455–74.

Fang, Di, Nayga R, Snell H, West G, and Bazzani C. 2019. Evaluating USA's New Nutrition and Supplement Facts Label: Evidence from a Non-Hypothetical Choice Experiment. *Journal of Consumer Policy* 42: 545–62.

Fiebig, Denzil G, Michael P Keane, Jordan Louviere, and Nada Wasi. 2010. The Generalized Multinomial Logit Model: Accounting for Scale and Coefficient Heterogeneity. *Marketing Science* 29: 393–421.

Fiore, Stephen M, Glenn W Harrison, Charles E Hughes, and Elisabet E Rutström. 2009. Virtual Experiments and Environmental Policy. *Journal of Environmental Economics and Management* 57(1): 65–86.

Fox, Jesse, Jeremy Bailenson, and Joseph Binney. 2009. Virtual Experiences, Physical Behaviors: The Effect of Presence on Imitation of an Eating Avatar. *Presence: Teleoperators and Virtual Environments* 18: 294–303.

Greene, William H, and David A Hensher. 2010. Does Scale Heterogeneity across Individuals Matter? An Empirical Assessment of Alternative Logit Models. *Transportation* 37: 413–28.

Hammond, David, Heather G Lillico, Lana Vanderlee, Christine M White, and Jessica L Reid. 2015. The Impact of Nutrition Labeling on Menus: A Naturalistic Cohort Study. *American Journal of Health Behavior* 39: 540–8.

Hensher, David A. 2010. Hypothetical Bias, Choice Experiments and Willingness to Pay. *Transportation Research Part B: Methodological* 44: 735–52.

Hess, Rebecca, Vivianne H Visschers, and Michael Siegrist. 2012. The Role of Health-Related, Motivational and Sociodemographic Aspects in Predicting Food Label Use: A Comprehensive Study. *Public Health Nutrition* 15: 407–14.

Howard, Gregory, Brian E Roe, Erik C Nisbet, and Jay F Martin. 2017. Hypothetical Bias Mitigation Techniques in Choice Experiments: Do Cheap Talk and Honesty Priming Effects Fade with Repeated Choices? *Journal of the Association of Environmental and Resource Economists* 4: 543–73.

Innocenti, Alessandro. 2017. Virtual Reality Experiments in Economics. *Journal of Behavioral and Experimental Economics* 69: 71–7.

Isacsson, Gunnar 2007. The Trade Off Between Time and Money: Is There a Difference Between Real and Hypothetical Choices? Working Paper 3, Swedish National Road & Transport Research Institute (VTI).

Jacquemet, Nicolas, Robert V Joule, Stephane Luchini, and Jason F Shogren. 2013. Preference Elicitation under Oath. *Journal of Environmental Economics and Management* 65(1): 110–32.

Kennedy, Robert S, Norman E Lane, Kevin S Berbaum, and Michael G Lilienthal. 1993. Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness. *International Journal of Aviation Psychology* 3: 203–20.

Lee, Jang Han, Jeonghun Ku, Kwanguk Kim, Byoungnyun Kim, In Y Kim, Byung-Hwan Yang, Seok Hyeon Kim, Brenda K Wiederhold, Mark D Wiederhold, Dong-Woo Park, Youngsik Lim, and Sun I Kim. 2003. Experimental Application of Virtual Reality for Nicotine Craving through Cue Exposure. *Cyber Psychology & Behavior* 6: 275–80.

Levy, Charles E, Shivashankar Halan, Erin P Silverman, Michael Marsiske, Leigh Lehman, David Omura, and Benjamin C Lok. 2015. Virtual Environments and Virtual Humans for Military Mild Traumatic Brain Injury and Posttraumatic Stress Disorder: An Emerging Concept. *American Journal of Physical Medicine & Rehabilitation* 94: e31–2.

Levy, Charles E, David M Miller, Christiana A Akande, Benjamin Lok, Michael Marsiske, and Shivashankar Halan. 2019. V-Mart, a Virtual Reality Grocery Store:

A Focus Group Study of a Promising Intervention for Mild Traumatic Brain Injury and Posttraumatic Stress Disorder. *American Journal of Physical Medicine & Rehabilitation* 98: 191–8.

List, John A, and Craig A Gallet. 2001. What Experimental Protocol Influence Disparities between Actual and Hypothetical Stated Values? *Environmental and Resource Economics* 20: 241–54.

Lusk, Jayson L, and F Bailey Norwood. 2009. An Inferred Valuation Method. *Land Economics* 85: 500–14.

Lusk, Jayson L, and Ted C Schroeder. 2004. Are Choice Experiments Incentive Compatible? A Test with Quality Differentiated Beef Steaks. *American Journal of Agricultural Economics* 86: 467–82.

Mahieu, Pierre-Alexandre, Pere Riera, and Marek Giergiczny. 2012. The Influence of Cheap Talk on Willingness-to-Pay Ranges: Some Empirical Evidence from a Contingent Valuation Study. *Journal of Environmental Planning and Management* 55: 753–63.

Matthews, Yvonne, Riccardo Scarpa, and Dan Marsh. 2017. Using Virtual Environments to Improve the Realism of Choice Experiments: A Case Study about Coastal Erosion Management. *Journal of Environmental Economics and Management* 81: 193–208.

Meißner, Martin, and Reinhold Decker. 2010. Eye-Tracking Information Processing in Choice-Based Conjoint Analysis. *International Journal of Market Research* 52: 593–612.

Murphy, James J, P Geoffrey Allen, Thomas H Stevens, and Darryl Weatherhead. 2005. A Meta-Analysis of Hypothetical Bias in Stated Preference Valuation. *Environmental and Resource Economics* 30: 313–25.

Nayga, Rodolfo M, Daria Lipinski, and Nitin Savur. 1998. Consumers' Use of Nutritional Labels while Food Shopping and at Home. *Journal of Consumer Affairs* 32: 106–20.

Norwood, F Bailey, and Jayson L Lusk. 2011. Social Desirability Bias in Real, Hypothetical, and Inferred Valuation Experiments. *American Journal of Agricultural Economics* 93: 528–34.

Papoutsi, Georgia S, Rodolfo M Nayga, Panagiotis Lazaridis, and Andreas C Drichoutis. 2015. Fat Tax, Subsidy or Both? The Role of Information and Children's Pester Power in Food Choice. *Journal of Economic Behavior & Organization* 117: 196–208.

Parsons, Thomas D. 2015. Virtual Reality for Enhanced Ecological Validity and Experimental Control in the Clinical, Affective and Social Neurosciences. *Frontiers in Human Neuroscience* 9: 660.

Penn, Jerrod M, and Wuyang Hu. 2018. Understanding Hypothetical Bias: An Enhanced Meta-Analysis. *American Journal of Agricultural Economics* 100: 1186–206.

Petrovici, Dan, Andrew Fearne, Rodolfo M Nayga, and Dimitris Drolias. 2012. Nutritional Knowledge, Nutritional Labels, and Health Claims on Food: A Study of Supermarket Shoppers in the South East of England. *British Food Journal* 114: 768–83.

Pieniak, Zuzanna, Wim Verbeke, and Joachim Scholderer. 2010. Health-Related Beliefs and Consumer Knowledge as Determinants of Fish Consumption. *Journal of Human Nutrition and Dietetics* 23: 480–8.

Sanchez-Vives, Maria V, and Mel Slater. 2005. From Presence to Consciousness through Virtual Reality. *Nature Reviews Neuroscience* 6: 332.

Sarrias, Mauricio, and Ricardo A Daziano. 2017. Multinomial Logit Models with Continuous and Discrete Individual Heterogeneity in R: the gmnl Package. *Journal of Statistical Software* 79: 1–46.

Savage, Scott J, and Donald M Waldman. 2008. Learning and Fatigue During Choice Experiments: A Comparison of Online and Mail Survey Modes. *Journal of Applied Econometrics* 23: 351–71.

Scarpa, Riccardo, and John M Rose. 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: 253–82.

Schulze, William D, Ralph C d'Arge, and David S Brookshire. 1981. Valuing Environmental Commodities: Some Recent Experiments. *Land Economics* 57: 151–72.

Silva, Andres, Rodolfo M Nayga, Jr, Benjamin L Campbell, and John L Park. 2011. Revisiting Cheap Talk with New Evidence from a Field Experiment. *Journal of Agricultural and Resource Economics* 36: 280–91.

Tyrrell, Ryan, Hilla Sarig-Bahat, Katrina Williams, Grace Williams, and Julia Treleaven. 2018. Simulator Sickness in Patients with Neck Pain and Vestibular

Pathology during Virtual Reality Tasks. *Virtual Reality* 22: 211–9.

US Food & Drug Administration. 2016. Changes to the Nutrition Facts Label. Available at: https://www.fda.gov/Food/GuidanceRegulation/GuidanceDocumentsRegulatoryInformation/LabelingNutrition/ucm385663.htm Accessed May 30, 2020.

Van Wezemael, Lynn, Vincenzina Caputo, Rodolfo M Nayga, George Chryssochoidis, and Wim Verbeke. 2014. European Consumer Preferences for Beef with Nutrition and Health Claims: A Multi-Country Investigation Using Discrete Choice Experiments. *Food Policy* 44: 167–76.

Vossler, Christian A, Maurice Doyon, and Daniel Rondeau. 2012. Truth in Consequentiality: Theory and Field Evidence on Discrete Choice Experiments. *American Economic Journal: Microeconomics* 4: 145–71.

Wu, Tina L, Adam Gomes, Keegan Fernandes, and David Wang. 2019. The Effect of Head Tracking on the Degree of Presence in Virtual Reality. *International Journal of Human–Computer Interaction* 35(17): 1569–1577.