International Journal of Research in Marketing 38 (2021) 70–84

Contents lists available at ScienceDirect

IJRM

International Journal of Research in Marketing

journal homepage: www.elsevier.com/locate/ijresmar

Full Length Article

A de-biased direct question approach to measuring consumers' willingness to pay☆

Reto Hofstetter [a], Klaus M. Miller [b,*], Harley Krohmer [c], Z. John Zhang [d]

[a] Faculty of Economics and Management, University of Lucerne, Frohburgstrasse 3, P.O. Box 4466, 6002 Lucerne, Switzerland
[b] Faculty of Economics and Business, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 4, 60629 Frankfurt am Main, Germany
[c] University of Berne, Engehaldenstrasse 4, 3012 Berne, Switzerland
[d] The Wharton School, University of Pennsylvania, 3730 Walnut St, Philadelphia, PA 19104, USA

ARTICLE INFO

Article history:
First received on May 13, 2019 and was under review for 5½ months
Available online 23 June 2020

Area Editor: Russell Winer

Keywords:
Market research
Pricing
Demand estimation
Direct estimation
Single question approach
Choice experiments
Willingness to pay
Hypothetical bias

Knowledge of consumers' willingness to pay (WTP) is a prerequisite to profitable price-setting. To gauge consumers' WTP, practitioners often rely on a direct single question approach in which consumers are asked to explicitly state their WTP for a product. Despite its popularity among practitioners, this approach has been found to suffer from hypothetical bias. In this paper, we propose a rigorous method that improves the accuracy of the direct single question approach. Specifically, we systematically assess the hypothetical biases associated with the direct single question approach and explore ways to de-bias it. Our results show that by using the de-biasing procedures we propose, we can generate a de-biased direct single question approach that is accurate enough to be useful for managerial decision-making. We validate this approach with two studies in this paper.

1. Introduction

The key to the optimal pricing decision for new and existing products and services is an accurate understanding of consumers' willingness to pay[1] (WTP; Anderson, Jain, & Chintagunta, 1992). Consumers' WTP is also important for implementing various pricing tactics, such as nonlinear pricing (Jedidi & Zhang, 2002), one-to-one pricing (Shaffer & Zhang, 2000), and targeted promotions

☆ We thank Bernd Skiera, Shunyao Yan, Rahul Thondan, and seminar participants at Goethe University Frankfurt, University of Geneva, Vienna University of Economics and Business, Bocconi University, and the University of Berne, as well as participants at the JPIM Research Forum, PSI, EMAC, and AMA conferences and two anonymous IJRM reviewers, the Area Editor Russ Winer, and the Co-Editor Werner Reinartz for helpful comments and feedback. This paper is based on the first two authors' dissertation, who are listed in alphabetical order and contributed equally to this paper. All remaining errors are our own.

* Corresponding author.
E-mail addresses: reto.hofstetter@unilu.ch, (R. Hofstetter), klmiller@wiwi.uni-frankfurt.de, (K.M. Miller), krohmer@imu.unibe.ch, (H. Krohmer), zjzhang@wharton.upenn.edu. (Z.J. Zhang).

[1] In this paper, we take the standard economic view of consumer willingness to pay and define it as the maximum price at or below which a consumer will definitely buy one unit of the product. This corresponds to the concept of the floor reservation price as proposed by Wang et al. (2007). However, we do not adopt their idea of conceptualizing WTP as a range. Instead, we consider WTP as a point measure, staying in line with earlier literature in economics on measuring consumer WTP (e.g., Miller et al., 2011; Wertenbroch & Skiera, 2002) and account for the individual variation of consumer WTP by constructing appropriate confidence intervals for our WTP measures at the aggregate level (see results section below). Further, we focus on the WTP for a product as a whole, assuming known availability and awareness of the product. Our study does not address the WTP for features of a product.

https://doi.org/10.1016/j.ijresmar.2020.04.006

Table 1
Alternative methods to measuring consumers' willingness to pay (WTP).

Context	Measurement			
	Direct		Indirect	
	Single question	Multiple question	Single question	Multiple question
Hypothetical WTP (HWTP)	e.g., Open-ended question format (OE)	e.g., Van-Westendorp Method (VWM)	e.g., Dichotomous-choice question format (DC)	e.g., Choice-based conjoint analysis (CBC)
Actual WTP (AWTP)	e.g., Becker, DeGroot, and Marschak mechanism (BDM)	e.g., Incentive-aligned measurement of WTP range (ICERANGE)	e.g., Incentive-aligned dichotomous-choice question format (IDC)	e.g., Incentive-aligned choice-based conjoint analysis (ICBC)
Hypothetical WTP (HWTP) De-biasing approaches	e.g., Consequentialism, honesty and realism approaches, cheap talk, uncertainty adjustment, Bayesian Truth Serum, response calibration		e.g., Dual-response choice designs, mental simulation, data fusion of hypothetical and actual WTP data	

Table 2
Comparison of alternative methods to measuring consumers' willingness to pay (WTP).

Measurement	Direct				Indirect			
Context	Hypothetical WTP		Actual WTP		Hypothetical WTP		Actual WTP	
Example methods	OE	VWM	BDM	ICERANGE	DC	CBC	IDC	ICBC
Quality of obtained information								
Measure	Stated preferences		Revealed preferences		Stated preferences		Revealed preferences	
Uncertainty on external validity	High		Low		High		Low	
Scope of information	WTP of entire product		WTP of entire product		WTP of entire product and product features[a]		WTP of entire product and product features[a]	
Capture real price sensitivity	+++		+++		++		+	
Ease of implementation								
Ease of data collection	++++		++		+++		+	
Ease of data analysis	++++		++++		++		++	
Interview time	++++		+++		++		+	
Required sample size	++++		+++		++		+	
Costs	++++		++		+++		+	
Applicability								
New products	Yes		No		Yes		No	
Existing products	Yes		Yes[b]		Yes		Yes[b]	

Notes: ++++ most favourable method; + least favourable method. OE = open-ended (OE) question format. VWM = Van Westendrop method. BDM = Becker, DeGroot, and Marschak mechanism. ICERANGE = incentive-aligned measurement of WTP range. DC = dichotomous-choice (DC) question format. CBC = choice-based conjoint analysis. IDC = incentive-aligned dichotomous choice question format. ICBC = incentive-aligned choice-based conjoint analysis.

[a] WTP for features of a product can only be obtained from indirect multiple question approaches such as choice-based conjoint analysis. We note that indirect single question approaches such as the dichotomous choice format or direct methods can be applied to individual features of a product, too. However, this would require a separate assessment of these individual features and subsequently more interview time as well as costs.

[b] At minimum, incentive-aligned methods require an existing product prototype that can be sold to the survey respondents.

(Shaffer & Zhang, 1995). Not surprisingly, various approaches have been developed to determine consumers' WTP (Miller, Hofstetter, Krohmer, & Zhang, 2011). The main distinctions among these approaches are whether they measure WTP directly or indirectly, whether they use a single question or multiple questions, and whether they determine consumers' actual or hypothetical WTP (see Table 1 for an overview of the various methods to measure WTP and Table 2 for a structured comparison of the advantages and disadvantages of each approach).

In practice, market researchers often ask consumers to state their WTP for a product directly (Anderson, Jain, & Chintagunta, 1992; Steiner & Hendus, 2012; Hofstetter, Miller, Krohmer, & Zhang, 2013). This can be done using either a single, open-ended (OE) question format (Arrow et al., 1993; Mitchell & Carson, 2013) or multiple, open-ended questions such as in the Van Westendrop Method (VWM; Van Westendorp, 1976). In a management survey we conducted of 82 pricing managers, we found that the direct approach is the most popular approach used to determine demand (used by 28%, see Web Appendix A). The direct approach is also widely used by market research firms. Steiner and Hendus (2012) find that 76% of the surveyed firms use a direct approach. The enduring popularity of the direct approach is due to its obvious advantages. Conceptually simple and easy to implement with regard to data collection and analysis, the direct approach unfailingly generates timely information at a low cost (Jedidi & Jagpal, 2009). Advances in digital technologies today make the direct approach seem to shine even more brightly because it facilitates massive online collections of consumer WTP data about a large number of products within a very short time.[2] The

[2] See also Brynjolfsson et al. (2019) who use an indirect, single question approach, the dichotomous-choice (DC) question format, to collect massive online choice experimental data to measure changes in well-being. In a similar vein, open-ended (OE) questions can easily be administered to a large number of respondents via the Internet.

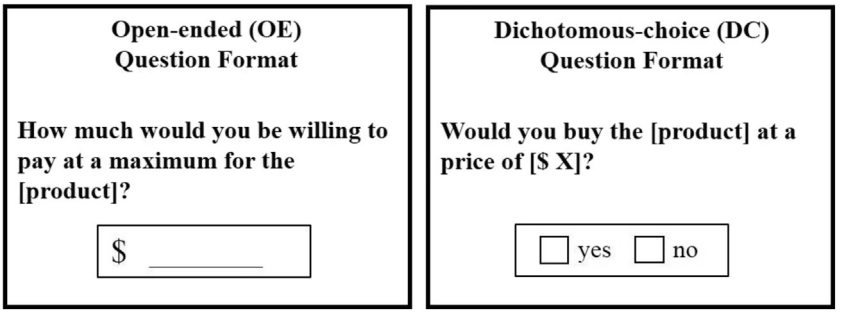


Fig. 1. Alternative single question formats to measure consumer willingness to pay (WTP).

popularity of the direct approach is also helped by its inclusion in commercial applications, e.g., the price sensitivity meter (PSM and PSM plus) from GfK and BASES Price Advisor from Nielsen.

An alternative way to measure WTP is to use the indirect approach, such as using a single dichotomous-choice (DC) question format (Mitchell & Carson, 2013), or multiple sequential questions such as in a choice-based conjoint analysis (CBC; Louviere & Woodworth, 1983). The indirect approach has been shown to capture more realistic choice and purchase scenarios (Leigh, MacKay, & Summers, 1984) and can, as in the case of conjoint analysis, provide additional information on the WTP for individual product attributes (e.g., Green & Krieger, 1996; Hanson & Martin, 1990; Jedidi & Zhang, 2002). The downside of these indirect approaches is that except perhaps for the DC question format, they all require more effort in data collection and more expertise in analysis. The effort and expertise required can be quite costly and discouraging for many practitioners, so much so that many shun those approaches.

One thing that both the direct and indirect approaches have in common is the fact that they all elicit consumers' hypothetical WTP (HWTP), because they do not typically require the respondents to actually buy the product (Ding, Grewal, & Liechty, 2005; Miller et al., 2011; Wertenbroch & Skiera, 2002). Hypothetical WTP, which corresponds to consumers' stated preferences, can deviate from their actual WTP (AWTP; Hoffman, Menkhaus, Chakravarti, Field, & Whipple, 1993), which reflects their revealed preferences. This deviation, known in the economics literature as hypothetical bias, is induced by the hypothetical nature of a task (Harrison & Rutström, 2008). One way to remove this bias is to use incentive-aligned direct methods to measure consumers' actual WTP. The BDM mechanism using a single question as proposed by Becker, DeGroot, & Marschak (BDM; 1964) is one such method, and the ICERANGE approach using multiple questions as suggested by Wang, Venkatesh, and Chatterjee (2007) is another. A consumer's actual WTP can also be measured indirectly using the incentive-aligned counterpart of the single dichotomous choice (DC) question format (Brynjolfsson, Collis, & Eggers, 2019) or multiple-question, incentive-aligned, choice-based conjoint analysis (ICBC; Ding et al., 2005; Ding, 2007; Dong, Ding, & Huber, 2010).[3]

For researchers and practitioners, an incentive-aligned approach should be the method of choice. However, an incentive-aligned approach may not always be feasible in the case where product prototypes are not available or privacy concerns and legal restrictions in some countries prevent their usage, or the incentives are too costly to provide or to simulate. Indeed, Europe's General Data Protection Regulation (GDPR) has rendered incentive-aligned methods that require the collection of respondents' personal data (e.g., email addresses, phone numbers, etc.) much more difficult and costly to implement. Even if incentive-aligned approaches are feasible, their application may be very costly, for example, for high-value items such as a car or a house. Further, incentive-aligned approaches put an excessive burden on the respondents in terms of understanding the research method as well as the required time to respond (Ding, 2007; Ding et al., 2005; Miller et al., 2011; Wertenbroch & Skiera, 2002).[4] For all these reasons, scholars and practitioners are looking hard for new ways to remove the hypothetical bias from non-incentive aligned approaches. Interestingly, in that pursuit, most scholars focus on the more complex, indirect approach—conjoint analysis—and have had a number of successes with that method (see e.g., Jedidi & Jagpal, 2009). These successes include formulas related to data calibration (Brazell et al., 2006), mental simulation (Hoeffler, 2003), and mixing hypothetical choice data with a small amount of incentive-aligned choice experimental data (Laghaie & Otter, 2019).

In contrast, academic researchers pay scant attention to simpler, direct approaches such as the OE question format, even though practitioners routinely use such an approach. In this paper, we will develop some rigorous de-biasing procedures for OE, the simplest direct approach in use. Our de-biasing strategy is to calibrate data from two single question formats, OE and DC, as illustrated in Fig. 1,[5] in a theory-informed way. Specifically, we systematically investigate the nature of hypothetical biases

[3] Other incentive-aligned approaches include a sequential incentive-compatible conjoint procedure for eliciting consumer WTP for attribute upgrades proposed by Park, Ding, and Rao (2008) or an incentive-compatible dynamic auction for selling multiple complementary goods as suggested by Sun and Yang (2014).

[4] We note that in practice, market researchers can also obtain consumers' actual WTP from real market transactions such as field experiments, scanner data, online purchases, or simulated test market data. Actual WTP data from past transactions is incentive compatible and shows a high convergent validity due to actual purchase observations under realistic market conditions. However, because consumers' true WTP remains unknown, the interpretation of these data is difficult (Wertenbroch & Skiera, 2002). Also, there is not always sufficient natural variation in prices or only within a very limited range for the focal firm's product and its competitive products to estimate the true WTP (Jedidi & Jagpal, 2009). Field experiments, most notably online these days, can provide some remedy, but similar to the various sources of transaction data mentioned above, they are not feasible for new products or are simply too expensive to implement.

[5] Some studies in the experimental economics literature refer to a third single question format, the payment-card (PC) question format (see Mitchell & Carson, 2013 for an overview). With the PC question format, respondents are asked to select one of several proposed prices as their maximum WTP for a specific product. If their maximum WTP differs from their proposed WTP, respondents can still give an individual value by using an open-ended field. We did not include this format in our study because it is in fact a hybrid question format that combines elements of price generation (OE) and price selection tasks (DC).

associated with the data from two single question formats in a marketing context based on past research and leverage the inherent bias structure associated with each question format to improve the data from OE. We show that by using the de-biasing procedures we propose here, the use of data from OE can help practitioners make more accurate managerial decisions without sacrificing the simplicity and timeliness they value (see Web Appendix I for a managerial guide on how to apply the de-biasing approach in market research practice). We demonstrate the value of our proposed de-biasing procedures through two online studies. In each study, our procedures perform remarkably well.

Previous studies have documented the existence of the hypothetical bias for various question formats (e.g., Balistreri, McClelland, Poe, & Schulze, 2001; Bishop, Welsh, & Heberlein, 1992; Harrison & Rutström, 2008) and the extent of this bias (Lusk & Schroeder, 2004; Miller et al., 2011; Schmidt & Bijmolt, 2020). A few researchers have looked specifically into which single question format yields the least biased (i.e., most valid) results by comparing two subsets of single question approaches, the DC and OE formats (Balistreri et al., 2001; Bishop et al., 1992; Brown, Champ, Bishop, & McCollum, 1996; Loomis, Brown, Lucero, & Peterson, 1997; Murphy, Geoffrey Allen, Stevens, & Weatherhead, 2005). A number of ex-ante and ex-post calibration techniques have been proposed for de-biasing single question-based pricing surveys (Carson, 2000; Loomis, 2011). Ex-ante techniques attempt to improve hypothetical methods at the data collection stage through priming survey subjects, whereas ex-post approaches try to calibrate data after measuring WTP. These techniques all rest on the assumption that, although responses to hypothetical questions may be biased, these responses provide useful information if extracted properly (Murphy & Stevens, 2004).

Ex-ante de-biasing techniques include letting survey subjects understand the consequences of their answers (Carson, Groves, & List, 2006; Cummings, Harrison, & Osborne, 1995; Landry & List, 2007), urging their honesty and realism before the survey (Jacquemet, Joule, Luchini, & Shogren, 2013; Loomis, González-Cabán, & Robin, 1996; Stevens, Tabatabaei, & Lass, 2013), or reminding them of possible biases (Cummings & Taylor, 1999; List, 2001; Poe, Clark, Rondeau, & Schulze, 2002; Aadland and Caplan, 2003; Brown, Ajzen, and Hrubes, 2003; Lusk, 2003). These prior studies have shown mixed results in eliminating the hypothetical bias, which is plausible given that consumers may not always adjust their actual behaviors based on verbal reminders alone (e.g., Farrell & Rabin, 1996).

Ex-post de-biasing approaches include methods such as uncertainty adjustment (e.g., Champ & Bishop, 2001; Champ, Bishop, Brown, & McCollum, 1997; Poe et al., 2002), the Bayesian Truth Serum (e.g., Prelec, 2004; Weaver & Prelec, 2013), and response calibration (e.g. Arrow et al., 1993; Fox, Shogren, Hayes, & Kliebenstein, 1998; Hofler & List, 2004; List & Shogren, 2002; Murphy et al., 2005; Murphy & Stevens, 2004). Ex-post de-biasing approaches can yield good approximations of actual payments in some applications, but not in others. It remains unclear, however, how well these ex-post de-biasing approaches would work in our specific context of measuring WTP for a consumer products, since they were developed in other contexts such as measuring WTP for non-market, public goods.

Our study differs from those cited above in three significant ways: First, we conduct our investigation in the context of pricing a regular consumer good that marketing practitioners typically deal with, not in the context of a public good. Second, we appeal to the past research to identify the theoretical bias structure associated with the data from each question format. Then, we collect single-source data to directly compare consumers' actual willingness to pay (AWTP) from BDM to the WTP information solicited through the single question format (i.e., OE and DC). This enables us to verify the theoretical bias structure of each single question format. Third, we base our de-biasing procedures on theory. We do so by analytically leveraging the bias structures inherent in each question format and identifying the ideal de-biasing procedure for the data elicited through OE. Based on that ideal procedure, we then propose three levels of precision in de-biasing for practitioners and validate our proposed procedures by applying them to the single-source data collected. In short, our research provides some practical de-biasing procedures for the direct question approach that are conceptually sound and complete with promising external validity.

It is important to note that our research here is not meant to promote the usage of the direct question approach or to pass judgment on the adequacy or inadequacy of alternative approaches such as conjoint analysis. Rather, it is to improve the accuracy of the direct question approach, if practitioners and researchers choose to use them for one reason or another.

2. Model development

In this section, we formalize a bias model for a price generation task (i.e., the OE question format) as well as for a price selection task (i.e., the DC question format). We will subsequently use these models to derive a robust de-biasing formula for estimating true WTP

Table 3
Model notation.

Notation	Meaning
p_i	Actual WTP for consumer i
$\tilde{p}_i$	Stated WTP for consumer i from an open-ended (OE) question format
$\hat{p}_i$	Stated WTP for consumer i from a dichotomous-choice (DC) question format
p_i^c	Price cue presented to consumer i in a dichotomous-choice (DC) question format
$\overline{p}$	Mean of actual WTP p_i
$\tilde{p}$	Mean of stated WTP $\tilde{p}_i$
$\hat{p}$	Mean of stated WTP $\hat{p}_i$
p^c	Mean of price cues p_i^c
α	Product category-level bias in an open-ended (OE) question format
ε_i	Individual-level bias for consumer i in an open-ended (OE) question format
θ_i	Individual-level bias for consumer i in a dichotomous-choice (DC) question format

based on data collected through OE and DC questions. See Table 3 for a summary of our model notation and definition of the respective variables.

Let p_i be the actual WTP for an individual consumer i where p_i is a random draw from $f(p_i)$ with $p_i \sim N(\bar{p}, \sigma^2_{p_i})$ and $p_i \geq 0$. Further, let $\tilde{p}_i$ be the stated WTP for an individual consumer i when stating her maximum WTP under an OE question format and $\hat{p}_i$ under a DC question format. Similarly, we define respective means as $\bar{p} = \frac{1}{n}\sum_{i=1}^{n} p_i$, $\tilde{p} = \frac{1}{n}\sum_{i=1}^{n} \tilde{p}_i$, and $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \hat{p}_i$.

For each respondent i in the two single question formats, we observe either $\tilde{p}_i$ or $\hat{p}_i$. If the respondent is unbiased, we should observe $p_i = \tilde{p}_i$ and $p_i = \hat{p}_i$, respectively. However, stated WTP often deviates from actual WTP, as consumers often need to construct their WTP in response to the specific elicitation context rather than retrieving a previously formed value stored in their memory (Bettman, Luce, & Payne, 1998; O'Donnell & Evers, 2019). Past literature has shown multiple reasons for both price generation tasks and price selection tasks to generate biases specific to question formats. In the following, we will draw on this literature to derive question format-specific bias models for price generation (OE) and price selection (DC) tasks, which we will use subsequently to de-bias the data from the OE format.

2.1. Biases in price generation formats (OE question format)

Price generation tasks, such as the OE question format, do not provide the respondent with any cues regarding reference prices or price ranges and hence allow respondents an unlimited degree of freedom to state their WTP (Chernev, 2003). There is little research on how consumers respond to price generation tasks (Chernev, 2006), however, the little we do know indicates that consumers have to go through a three-stage price generation process. First, they need to evoke the range of possible values. Next, they use these values as benchmarks to determine the product's potential utility to themselves and articulate this utility on a monetary scale, thus forming their maximum WTP (Chernev, 2003). Finally, when consumers state their true WTP at the moment of the survey, they need to discount their hypothetical future WTP (Loomis et al., 1997). As the effort in this process is generally not sufficiently rewarded (Arrow et al., 1993; Kemp & Maxwell, 1993), respondents may not properly discount their WTP and hence report inflated values leading to an individual, consumer-level bias (Ding et al., 2005; List, 2001; Miller et al., 2011). This inflated individual-specific bias is also confirmed by a study from Loomis et al. (1997). In that study, respondents typically inflate the amount they would pay in a hypothetical purchasing situation as compared to an incentive-aligned purchasing situation. This lack of incentive-compatibility leading to an overstated WTP for price generation tasks is also noted by Hoehn and Randall (1987) and Carson, Groves, and Machina (2000). Therefore, we can write stated WTP $\tilde{p}_i$ from OE as

$$\tilde{p}_i = (\alpha + \varepsilon_i) + p_i. \tag{1}$$

In other words, a consumer's stated WTP is an inflated value of her true WTP and the inflation factor $\alpha + \varepsilon_i$ is individual-specific. Here, $\varepsilon_i \sim N(0, \sigma^2_{\varepsilon i})$ captures individual-specific variations and α is a product category-specific inflator that is typically positive (but is not required to be) as noted in Dickie, Fisher, and Gerking (1987), Shogren (1990), Carson et al. (1996), as well as List and Shogren (1998). In summary, Eq. (1) describes the well-documented bias that an individual tends to inflate her true WTP with an inflator that varies by the individual, but will have a category-specific positive bias.

2.2. Biases in price selection formats (DC question format)

Price selection tasks, such as the dichotomous-choice (DC) question format, lighten the burden of setting a monetary value for a product by presenting the consumer with a single price for the product under study. The consumer's decision in this case is reduced to whether to select the product at the given price.

It is well known that the price selection task has its own issues. According to Kahneman, Slovic, and Tversky (1982), people under uncertainties are prone, when making estimates, to start from an initial value, which they will then adjust to yield the final answer. Such an anchoring strategy will bias toward the initial value because the adjustments are typically insufficient (Slovic & Lichtenstein, 1971). In price selection tasks, anchoring occurs when respondents seize upon the price cue as the value of the product (Mitchell & Carson, 2013). Studies have shown that such anchoring is indeed common. The respondent confronted with a dollar figure in a situation where she is uncertain about a product's value, the respondent may regard the proposed amount as conveying approximation of the product's true value. She will then anchor her WTP on the proposed amount (Green, Jacowitz, Kahneman, & McFadden, 1998; Mitchell & Carson, 2013). That is, respondents have a propensity to inflate their true WTP when faced with a high price cue. Conversely, when confronted with a low price cue, respondents tend to underreport their true WTP.

This tendency of anchoring to inflate or deflate actual WTP is also rooted in the perception that price and quality are often correlated, to the extent that a price cue may be taken as an indicator of product quality (Gabor & Granger, 1966; Monroe, 1973; Shapiro, 1968). According to the literature on the price-quality relationship, consumers who make price-quality inferences tend to prefer higher-priced products when price is the only information available, as they apparently perceive the more expensive product to be of higher quality (John, Scott, & Bettman, 1986; Monroe, 1973). Consumers in this situation may be more likely to accept prices greater than their actual WTP. This may explain so-called *yea-saying* (i.e., consumers accept a given price as their WTP although their actual WTP is lower), which has been observed in experimental economics (e.g., Brown et al., 1996; Mitchell & Carson, 2013).

In contrast, if a relatively low price is presented to the consumer, she may perceive the product to be of low quality. In this case, consumers will be less likely to accept prices lower than their actual WTP. This phenomenon is known as *nay-saying*, that is, consumers reject a given price, although their actual WTP is greater than that amount (Carson, 2000). Both yea- and nay-saying is more likely to occur the farther away the price cues are from a respondent's actual WTP. These biases are likely smaller when the price cues are closer to her actual WTP (Bishop et al., 2017; Carson, 2000; Carson et al., 2000; Green et al., 1998; Mitchell & Carson, 2013). The previous discussions suggest that we can specify a consumer's stated WTP elicited through the dichotomous-choice format (DC) as:

$$\hat{p}_i = \theta_i(p_i^c - p_i) + p_i \tag{2}$$

where θ_i with a zero mean is the individual's specific bias under a dichotomous-choice (DC) format and p_i^c is the price cue presented to consumer *i*. Eq. (2) captures the fact that under DC, a respondent's bias is anchored on the price cue presented to her; a higher (lower) price cue than her actual WTP will lead her to state an inflated (deflated) WTP.[6]

2.3. De-biasing approach

As stated in the introduction, both OE and DC are easy to implement in practice. But once both data series are collected, the question becomes, how can we use these data to make managerial decisions? In the previous discussion, we showed that neither data series is ideal because of its inherent biases. However, we show in this section that by leveraging the bias structures of these two data series, we can actually de-bias the data series elicited through the OE question format so that it becomes possible to make all relevant managerial decisions unhampered by data biases.

To see how we can de-bias the OE series, we note that from Eq. (1), we have $\frac{1}{n}\sum_{i=1}^{n}\tilde{p}_i = \alpha + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i + \frac{1}{n}\sum_{i=1}^{n}p_i$. Given the zero mean for ε and definitions in Table 3, this implies

$$\alpha = \tilde{p} - \overline{p}. \tag{3}$$

Then from Eq. (2), once again using the definitions in Table 3, we have $\hat{p} = \frac{1}{n}\sum_{i=1}^{n}\theta_i p_i^c - \frac{1}{n}\sum_{i=1}^{n}\theta_i p_i + \overline{p}$. As price cues are randomly assigned to respondents, p_i^c and θ_i are independent variables.[7] This implies $\frac{1}{n}\sum_{i=1}^{n}\theta_i p_i^c = 0$, and $\hat{p} = \overline{p} - \frac{1}{n}\sum_{i=1}^{n}\theta_i p_i$. Also, note $\frac{1}{n}\sum\theta_i p_i = cov(\theta, p)$ (see Web Appendix C), where $cov(\theta,p)$ is the covariance. We must have

$$\overline{p} = \hat{p} + cov(\theta, p). \tag{4}$$

From Eqs. (3) and (4), we arrive at

$$\alpha = \tilde{p} - \hat{p} - cov(\theta, p). \tag{5}$$

Finally, from Eqs. (1) and (5), we can derive the bias correcting function below

$$p_i = \tilde{p}_i - \tilde{p} + \hat{p} + cov(\theta, p) - \varepsilon_i. \tag{6}$$

Eq. (6) suggests that a researcher can derive a respondent's actual WTP from stated WTP elicited from the OE question format by adding three non-individual, specific adjustment factors: $\tilde{p}$, $\hat{p}$, and $cov(\theta,p)$ plus ε_i, a white noise. The first two factors are known from the two biased data series collected through OE and DC and the third factor can be simulated easily as it is a constant number.

This enabled us to conduct our de-biasing procedure for the OE question format in four steps. First, we collected the two biased data series. Second, we de-biased the OE series by using $\tilde{p}$ and $\hat{p}$ only in Eq. (6). In other words, we set $cov(\theta,p) = 0$ and ignore ε_i in Eq. (6) as the first order approximation. We call this our BASIC de-biasing procedure. Third, we introduced ε_i while still keeping $cov(\theta,p) = 0$. We refer to this as our EPSILON de-biasing procedure. Finally, we introduced both $cov(\theta,p) \neq 0$ and ε_i. We call this our FULL de-biasing procedure. By following these four steps, we could tease out how effective each part of our de-biasing procedure for the OE data series is in helping us to make better managerial decisions. Finally, we check the results of these de-biasing procedures by measuring them against the gold standard of the consumer's actual WTP measurement, the output of the BDM mechanism (Becker et al., 1964; Miller et al., 2011; Wertenbroch & Skiera, 2002).

It is important to note that these proposed de-biasing procedures do not presume any knowledge of true WTP; we have collected the true WTP data through the BDM mechanism here merely to validate our procedures.

[6] At the zero mean θ_i can still have a distribution where there are more incidences for positive values than for negative values, such that a respondent is more likely to be biased upward (downward) at a high (low) price cue (see Web Appendix B).

[7] Random assignment is a sufficient condition, but not a necessary one. The necessary and sufficient condition is p_i^c is not assigned conditional specifically on θ_i.

3. Method and data collection

We conducted two large-scale empirical studies to test our de-biasing procedures. These studies gauge WTP for two new products, a gym bag and a sweatshirt, among students at a major Swiss university. Both products were specifically produced for the purpose of this study and new to the target market at the time of the study. We report the gym bag study (Study 1) here in detail and give a summary of the sweatshirt study (Study 2).

3.1. Participants

The data for the gym bag study was collected through an online experiment. To recruit participants, we sent out 12,448 invitation emails to the entire student body (undergraduate, graduate, and Ph.D. candidates) of a large Swiss university. We motivated participation by offering all survey participants a chance to win an Apple iPhone 7 Plus in a raffle[8]. The participants were further informed that their chance to win the raffle was independent of their experimental responses[9]. A total of 826 participants chose to take part in the bag study, which represents a response rate of 6.64%. We pre-specified that data collection would end after seven days (i.e., the decision to stop collecting data was independent of the experimental results; we did not analyze the data until after data collection had been completed). Within the seven-day period, we collected as much data as we could.

3.2. Stimulus

The stimulus we used in our study was a gym bag imprinted with a logo of the university that was not available in the market at the time of the study (see Web Appendix D for a depiction of the stimulus). The gym bag was designed and fabricated exclusively for the purpose of the study. We expected the university gym bag to be both interesting and affordable for most of the students, who actually represent the target segment of the product. Since apparel and accessories are also often sold online, the online channel represents the known and accustomed distribution channel for these categories[10]. Further, because the students had no reference for the exact market price for this distinctive university gym bag, their hypothetical WTP or actual WTP statements would not be capped[11].

As our stimulus was new to the market, no repeat purchases were observed and participants had never stated their WTP for the product before. Further, the university gym bag was not displayed in a competitive setting. As a result, students were unable to select the stimulus from a group of competing products as they would in a real online store. Finally, because we were conducting an online experiment, shipping the product had to be easy and cheap, which was the case with the gym bag.

3.3. Experimental design

We developed three different independent experimental groups and used a between-subjects design. Each participant was randomly assigned to one treatment group.

In the open-ended (OE) question format group, each participant directly stated her hypothetical WTP for the university gym bag.

In the dichotomous-choice (DC) question format group, we used a total of 21 price levels, ranging from Swiss Franc (CHF[12]) 1.25 up to CHF 26.25 incremented in steps of CHF 1.25. We chose the market price of the most expensive gym bag similar to our products as our upper limit of CHF 26.25. Each respondent received a price level that was chosen randomly from the 21 available levels. The random distribution of the price levels was even, meaning that all the price levels had an equal 1 in 21 chance of appearing in the respondent's DC question.[13]

In the BDM group, which is our control group for validating our de-biasing procedures, we determined our benchmarking actual WTP data by using an incentive-aligned mechanism, the BDM mechanism proposed by Becker et al. (1964). We chose the BDM mechanism as WTP from BDM has been found to not significantly differ from a consumer's WTP based on real purchase data (Miller et al., 2011). We implemented the BDM mechanism in a way similar to what Wertenbroch and Skiera (2002) did. In our particular application of the BDM mechanism, we told participants that they would have a chance to purchase the university gym bag without having to invest more money than they would be willing to pay for the product. We also informed them that the price for the university gym bag was not yet set, and that it would be determined randomly from a predefined uniform distribution unknown to the participants. Participants were further told that they were obligated to buy the university gym bag at the

[8] We used the smartphone as a single incentive to motivate participation in our survey in order to recruit an adequate number of subjects. However, the smartphone was not connected to our stimulus and the incentive-aligned condition under BDM, where proper incentives are offered to the participants so that they are motivated to reveal their true preferences (see Wertenbroch & Skiera, 2002 for details).

[9] It is possible that some consumers may have taken part in the survey just to win the smartphone and may not have been interested in purchasing the gym bag.

[10] See e.g., the Stanford Bookstore online: https://www.bkstr.com/stanfordstore/home.

[11] We acknowledge, however, that some participants may have a reference price from similar products in mind.

[12] At the time when this paper was written 1 CHF represents approximately 1 USD.

[13] In Web Appendix H, we test the robustness of our de-biasing approaches to the chosen DC price range in a simulation study. We find that the range of prices chosen in DC matters. If the DC range chosen is larger than the range of true WTP, reducing the DC range will not have any impact on our de-biasing procedures. However, if the DC range is contained by the true WTP, our de-biasing procedure is still robust, if we do not reduce the DC range by >35% of the true price range. Beyond that, our de-biasing procedures do not work as well and some work better than others. The de-biasing procedures do not work as well because an unrealistic range of prices in DC will simply distort the estimation of mean WTP. The simulation shows that the DC range needs to be carefully chosen to reflect the true minimum and maximum WTP in the market and that it is better to err on the wide side than on the narrow side of the DC range. In addition, we show that a parametric approach can help reduce biases if the DC range chosen is too narrow. We find the parametric approach to be highly robust to misspecifications of the DC range.

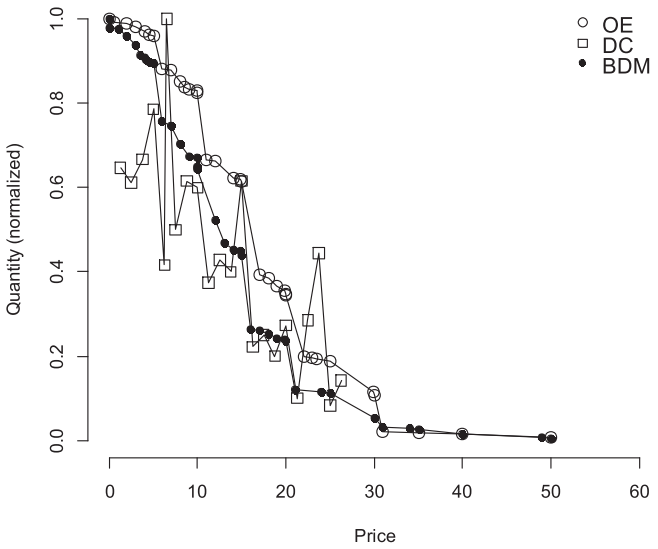


Fig. 2. Aggregate demand curves for university gym bag based on different WTP measurement approaches. Note: The curves show the aggregate demand based on aggregating individuals' responses to the OE, DC, or BDM questions. OE: open-ended question format; DC: dichotomous-choice question format; BDM: Becker, DeGroot, and Marschak mechanism; Quantity is the aggregate demand that is normalized to a range between 0 and 1 due to differing sample sizes. Prices are in CHF (Swiss Francs), which at the time of the study equal approximately to prices in USD.

randomly determined price if the price was less than or equal to their stated WTP. However, if the randomly determined price was higher, a respondent would not have a chance to buy the product. This mechanism ensures that participants have no incentive to state a price that is higher or lower than their true WTP.

To carry out the buying obligation, we recorded the name and address of each participant in the BDM group. After the completion of the study, we determined each individual participant's buying obligation by drawing from a discrete uniform distribution which corresponded to the price-levels used under the DC question format. The distribution thus ranged from CHF 1.25 to CHF 26.25 incremented in steps of CHF 1.25 and included a total of 21 price levels. We determined per participant whether the randomly drawn price was smaller or equal to her stated WTP. Thus, none of the participants had to purchase the gym bag at a price that was larger than their stated WTP. Out of all participants in the BDM group, 19.85% were obliged to buy the product at an average price of 8.03 ($SD = 5.66$, $min = 1.25$, $max = 26.25$). Only 12 participants (21.81%) of those who were obliged to buy paid a price higher than the average BDM WTP of 13.04. After the completion of the study, all participants who were obligated to buy were sent the gym bag with an invoice, via the post. The invoice was due within 14 days and payable with cash or credit card (this payment process was officially approved by the appropriate university authorities). Out of all 277 respondents in the BDM group, 55 (19.86%) were required to purchase the bag. Only one respondent refused to comply with her purchase obligation and returned the product[14].

We obtained 270 responses in the OE group, 279 responses in the DC group, and 277 responses in the BDM group. Our realized sample size exceeds current expectations in experimental studies of larger than 50 respondents per cell (Simmons, 2014).

The three experimental groups did not differ significantly in terms of socio-demographics or socioeconomic status. We performed a multivariate analysis of variance (Pillai-Spur: $F = 0.922$, $p = .562$) for age ($F = 0.465$, $p = .707$), sex ($F = 2.077$, $p = .102$), education ($F = 0.139$, $p = .937$), occupation ($F = 0.593$, $p = .620$), income ($F = 2.028$, $p = .109$), budget for clothing and accessories ($F = 0.1886$, $p = .904$), and purchase interest ($F = 0.604$, $p = .613$).

3.4. Experimental procedure

We divided our online experiment into three parts. The first part described the product (i.e., the university gym bag) in the OE, DC, and BDM groups. The second part consisted of the WTP task in the different experimental treatment groups. In the third part of the online experiment, we conducted a brief survey on socio-demographics and -economics, and we made sure the participants understood the WTP elicitation method to which they were exposed (see Web Appendix E for further details).

[14] Valid actual WTP estimation requires that the respondents understand the BDM procedure and the underlying buying process. In our sample, respondents understood the BDM mechanism quite well. As Wertenbroch and Skiera (2002) and Miller et al. (2011) did, we asked the subjects if it was clear to them why it was in their best interest to state exactly the price they were willing to pay. Using a seven-point Likert scale from one (not at all) to seven (very much so), the participants responded with an average of 5.954. We used a similar method to determine the understanding of the buying process and found an average of 6.222. Finally, we asked respondents if stating their WTP for the product was a task which was easy to understand and complete and participants replied with an average of 5.851. (see Web Appendix E for the exact wording).

3.5. WTP estimation procedure

Fig. 2 plots the observed demand in each treatment group. For the OE and BDM groups, we obtained each respondent's hypothetical WTP and actual WTP directly from the survey data and plotted demand $q(p)$ as the probability that a respondent's WTP is equal to or greater than a certain price p using the demand function of the form $q(p) = Pr\,(WTP \geq p)$. For the DC group, we plotted the choice share for each price level. We determined the face validity of WTP measures by correlating elicited WTP with the respondent's purchase interest. We measured purchase interest itself by using a seven-point Likert scale ranging from one (low interest in the product) to seven (high interest in the product). Face validity is high for all methods because correlations are positive and significant (OE: $r = 0.390$, $p < .001$, BDM: $r = 0.216$, $p = .001$). We did not test the DC question format because hypothetical WTP data was not available on an individual level.

4. Results

4.1. Study 1: university gym bag

First, we eyeball the empirical demand functions that can be generated from the three experimental groups (Fig. 2). We observe that OE demand (circles) is systematically higher compared to BDM (dots) at most price levels. For DC (squares), the function appears to be underestimating demand at low prices and overestimating demand at high prices. Both observations are consistent with the biases identified in the literature, which we discussed in the Model development section.

Next, we compare statistically both OE and DC data series to the BDM. As summarized in Table 4, we find the OE mean to be statistically different from BDM as indicated by a *t*-test [$t(543.92)_{\text{OE vs. BDM}} = 4.167$, $p < .001$] and non-overlapping 95%-confidence intervals. It is also different from BDM in terms of the distribution of the data series, as indicated by a KS-test and LR-test ($D_{\text{KS-Test, OE vs. BDM}} = 0.181$, $p < .001$; $D_{\text{LR-Test, OE vs. BDM}} = 41.615$, $p < .001$). However, the DC mean does not statistically differ from the BDM as indicated by overlapping 95%-confidence intervals, as we expected, but does differ in distribution ($D_{\text{LR-Test, DC vs. BDM}} = 18.383$, $p < .001$). Thus, our statistical analysis supports our model specifications for the biases identified in the literature for OE as well as DC data series.

We now show, in two steps, that by using our de-biasing procedure on the OE data series, we can, firstly, significantly improve its statistical fit with BDM and, secondly, demonstrate that the de-biased data series can help managers make better managerial decisions. Table 4 summarizes the results of our first step. In the Table 4, M_{BASIC} is the mean of the data series we generated by applying the BASIC de-biasing procedure, i.e., subtracting $\tilde{p}-\hat{p}$ from $\tilde{p}_i$ while setting $cov(\theta,p) = 0$ and $\varepsilon_i = 0$ (see Eq. (6)). This is the most straightforward case of de-biasing. We also generated M_{EPSILON}, the mean of the data series where we subtract $\tilde{p}-\hat{p}$ from $\tilde{p}_i$ while setting $cov(\theta,p) = 0$ and adding ε_i. In this simulation, we randomly drew ε_i from a normal distribution with zero mean and the same standard deviation as the OE data series $SD(\tilde{p}_i)$. Finally, M_{FULL} is the mean of the data series we have generated by fully simulating Eq. (6). Here, $cov(\theta,p)$ is simulated in the wide range of $[-3.08, 7.08]$, and in that range, the best theoretical ($cov(\theta,p) = 2.08$) and empirical ($cov(\theta,p) = 2.33$) outcomes for M_{FULL} are reported in Table 4 (Later in the paper, we will elaborate on how to select this range of $cov(\theta,p)$ and determine its theoretical best value.)

From Table 4, we can see that our BASIC de-biasing procedure has generated a data series that is closer to the BDM data series. Unlike the original OE data series, the 95%-confidence interval of the mean from the de-biased data series now overlaps with that of the BDM, although the *t*-test [$t(543.94)_{\text{BASIC vs. BDM}} = 2.76$, $p < .01$] and the KS-test ($D_{\text{KS-Test, BASIC vs. BDM}} = 0.26$, $p < .01$) remain negative. When individual-specific variations are accounted for in generating M_{EPSILON}, we generate a data series that is even closer to the BDM. Table 4 shows that this data series differs from the BDM only in distribution as indicated by the negative KS-test ($D_{\text{KS-Test, EPSILON vs. BDM}} = 0.22$, $p < .01$), but not the mean [$t(513.5)_{\text{EPSILON vs. BDM}} = 1.20$, $p > .05$]. Finally, when both covariance and individual-specific variations are incorporated in the FULL approach, we similarly find that this data series differs from the BDM

Table 4
Statistical analysis of collected data and de-biased data of university gym bag demand.

Data source	Mean [confidence interval]
Collected data	
OE	16.046[a, b, c] [15.041, 17.051]
DC	10.954[b] [10.402, 12.271]
BDM	13.041 [12.037, 14.044]
De-biased data	
$\text{BASIC}_{[\text{cov}=0,\ \text{epsilon}=0]}$	11.072[a, b] [10.089, 12.055]
$\text{EPSILON}_{[\text{cov}=0,\ \text{epsilon}=\text{SD(OE)}]}$	12.093[b] [10.851, 13.335]
$\text{FULL}_{[\text{cov}=2.08,\ \text{epsilon}=\text{SD(OE)}]}$	13.854[b] [12.536, 15.172]
$\text{FULL}_{[\text{cov}=2.33,\ \text{epsilon}=\text{SD(OE)}]}$	13.916[b] [12.580, 15.256]
BDM	13.041 [12.037, 14.044]

Note: OE: open-ended question format; DC: dichotomous-choice question format; BDM: Becker, DeGroot, Marschak mechanism; BASIC, EPSILON, FULL: refer to the steps of our de-biasing procedure; Values are shown with their 95% confidence interval in brackets; Superscript a indicates significant difference in terms of mean (*t*-test) relative to the benchmark; Superscript b indicates significant difference in terms of distribution (KS-test). For the DC data we used a Likelihood ratio test for the distribution comparison and compared confidence intervals (calculated based on Krinsky and Robb's (1986) procedure) for the mean comparison; Superscript c indicates non-overlapping confidence intervals relative to the BDM benchmark.

Table 5
Economic analysis results for university gym bag.

Data source	*Optimal price*	*Optimal quantity*	*Optimal profit*	*Profit percentage difference to BDM*
OE	15.000 [13.960, 15.990]	0.615[c, d] [0.546, 0.684]	6,148.148[c, d] [5,557, 6,737]	38.22%[c,d]
DC	19.340[c, d] [15.500, 25.470]	0.284[c, d] [0.230, 0.337]	4,024.250 [3,056, 5,315]	−9.52%
$BASIC_{[cov = 0, epsilon = 0]}$	14.810 [13.780, 15.830]	0.360 [0.270, 0.440]	3,487.110[d] [2,916, 4,021]	−21.60%[d]
$EPSILON_{[cov = 0, epsilon = SD(OE)]}$	15.748 [12.590, 17.860]	0.344 [0.276, 0.427]	3,702.102 [3,054, 4,260]	−16.77%
$FULL_{[cov = 2.08, epsilon = SD(OE)]}$	16.070 [11.470, 19.280]	0.378 [0.285, 0.489]	4,182.080 [3,447, 4,709]	−5.97%
$FULL_{[cov = 2.33, epsilon = SD(OE)]}$	15.425 [7.910, 18.130]	0.419 [0.356, 0.609]	4,362.976 [3,560, 4,929]	−1.91%
BDM	14.900 [14.560, 15.190]	0.449 [0.389, 0.510]	4,447.826 [3,865, 5,012]	N.A.

Note: OE: open-ended question format; DC: dichotomous-choice question format; BASIC, EPSILON, FULL: refer to the steps of our de-biasing procedure; BDM: Becker, DeGroot, Marschak mechanism; Quantity scaled from [0, 1]; N.A. = not applicable; Values are shown with their 95% confidence interval in brackets. We checked for a non-overlapping of confidence intervals as a test for significant differences; Superscript c indicates non-overlapping confidence intervals; Superscript d indicates a significant difference at $p < .05$ (bootstrapping the difference between the measures); The shaded cells indicate that the confidence interval of the specific measure overlaps with the confidence interval of the corresponding benchmark measure obtained from our BDM data. Thus, shaded areas imply no statistical difference between the estimated measure and the benchmark.

only in distribution as indicated by the negative KS-test when using our theoretically best value for $cov(\theta,p) = 2.08$ ($[t(505.31)_{FULL\ (cov=2.08)\ vs.\ BDM} = 0.97, p > .05]$, $D_{KS\text{-}Test,\ FULL(cov=2.08)\ vs.\ BDM} = 0.14$, $p < .05$, as well as overlapping 95%-confidence intervals) and for our empirically best value for $cov(\theta,p) = 2.33$ ($[t(502.03)_{FULL(cov=2.33)\ vs.\ BDM} = 1.03, p > .05]$, $D_{KS\text{-}Test,\ FULL(cov=2.33)\ vs.\ BDM} = 0.16, p < .01$, as well as overlapping 95%-confidence intervals). Clearly, our de-biasing procedures have generated a data series that is statistically indistinguishable from the BDM as far as these two tests are concerned.

Of course, whether a de-biased data series is a winner will depend on how well it can help a firm make its managerial decisions. Here, we can use the optimal price, optimal quantity, and optimal profits generated from the BDM data as the benchmarks and then compare them to the same variables generated from the original OE and DC data as well as from the three de-biased data series discussed above (BASIC, EPSILON, and FULL). For instance, the BDM data series gives us the incentive-compatible relationships between price and sales quantity. For the manufacturer of the university gym bag, the profit function is $\pi = (p - c) \times q(p) \times ms$, where p is the price, c are the marginal costs, q(p) is quantity scaled from [0,1] given by $q(p) = e^{\alpha+\beta\times p} \Big/ 1+e^{\alpha+\beta\times p}$ and ms is the market size. By incorporating the marginal cost for the university gym bag, which is c = CHF 5.00[15], and ms = 1000 as the size of the target market, we can easily derive the optimal price, optimal quantity, and optimal profits[16]. Note that for the purpose of optimization, a nonzero fixed cost will not alter the outcome. Similarly, we can use the same cost and market size information for all other data series to generate the equivalent numbers, which we can then compare with those generated with the BDM data series. To make statistical comparisons, we use two measurements: overlapping confidence intervals for each variable comparison and the variable difference test.

To generate the confidence interval for an optimal variable from a data series, such as price, we bootstrap a data series 1000 times and each time we generate an optimal price so that optimal prices from bootstrapping the data series generates a distribution for the optimal price (Efron & Tibshirani, 1993). This way, we can compare the 95%-confidence interval from different data series to see if they overlap. If they do not, we conclude that the optimal prices from the two data series are different.

To do the variable difference test, we once again bootstrap the two respective data series 1000 times and each time we calculate the difference in an optimal variable from two data series (Efron & Tibshirani, 1993). Thus, we generate a distribution of the differences of the two optimal variables. We then test if the mean difference in the optimal variable between the two bootstrapped data series is zero. If not, we conclude that the two optimal variables are statistically different.

In Table 5, we summarize the results of this analysis (for a visualization of the different demand curves resulting from the de-biasing approaches see Web Appendix Fig. F.1). Relative to the outcomes using the BDM data, a firm would significantly inflate its estimates of the optimal quantity when using the original OE data, while the optimal price is higher but passes both tests. The consequence is that the firm's profit estimate based on OE data would be 38.22% higher than the estimate from the BDM data. This inflation can have dire consequences in the firm's decision making. The estimates from the DC data also show considerable biases in a much higher optimal price and a much lower optimal quantity, so that both variables fail the two tests. However, because of the compensating nature in the inflated optimal price and deflated optimal quantity, the estimate of the optimal profit would not be significantly different from the estimate of the BDM data by both tests. Of course, even if the biases cancel each

[15] According to the manufacturer, variable costs did not depend on the number of units produced. However, variable costs may only be constant over a certain range around the actual quantity ordered by the manufacturer.

[16] Since the university gym bag was not available for purchase elsewhere at the time of the study and due to the university branding did not have any direct competitors, we assumed a monopoly for the purpose of price optimization. In order to consider indirect competitors (e.g., other gym bag brands), researchers would need to estimate demand not only for one product, but also for all competitive products and allow for competitive equilibrium analysis.

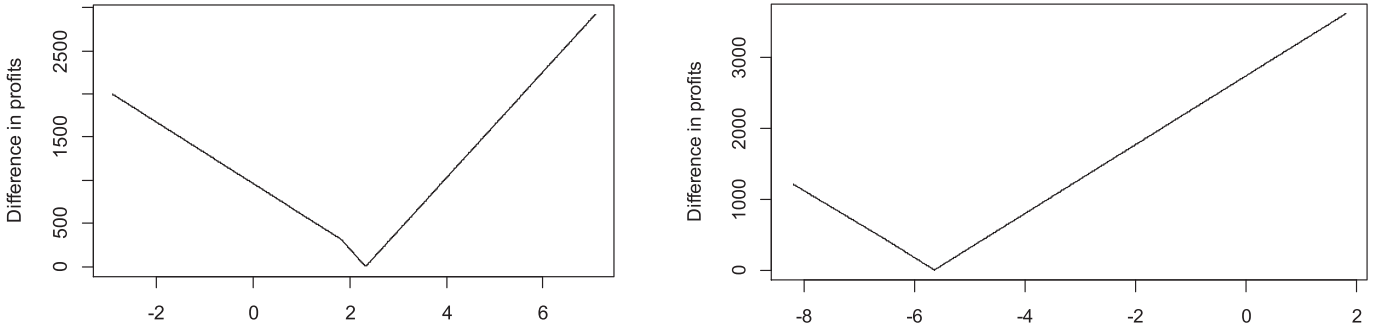
Study 1: University Gym Bag

$cov(\theta, p) \epsilon [-3.08, 7.08]$

Study 2: University Sweatshirt

$cov(\theta, p) \epsilon [-8.20, 1.80]$

Difference in profits across different values of cov(θ, p) when $SD(\varepsilon_i) = 0$



Difference in profits across different values of cov(θ, p) when $SD(\varepsilon_i) = SD(\tilde{p}_i)$

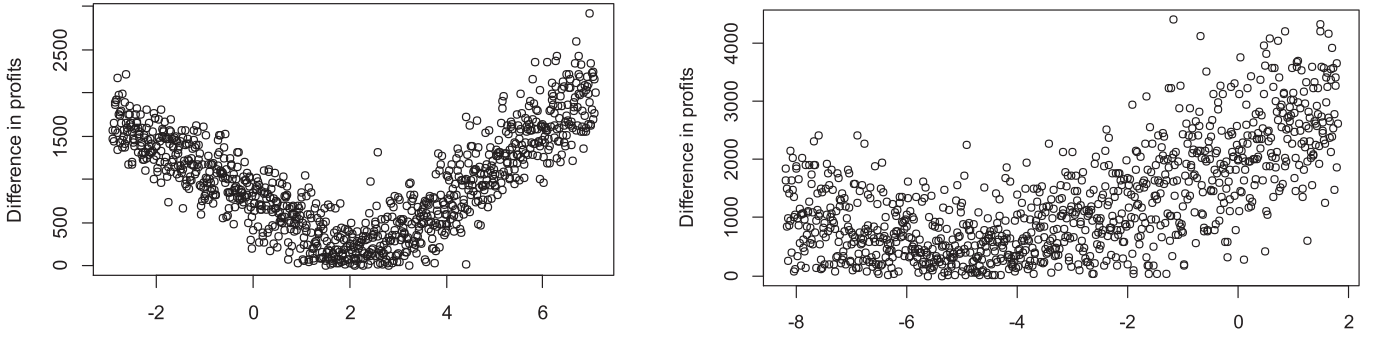


Fig. 3. Sensitivity analysis for differences in profits by varying COV (***θ***, p).

other out to a degree and generate a good estimate of profitability, the use of the DC data can still have severe consequences for a firm in production and marketing decisions.

From Table 5, we can see that the two partial de-biasing procedures (BASIC and EPSILON) have both improved the estimates of the optimal price and optimal quantity, and both estimates pass the two tests. Relative to the original OE data, both data series have significantly improved the estimate of the optimal profits in absolute difference. Most importantly, with the FULL de-biasing procedure, our estimates of the optimal price, quantity, and profits all pass the two tests and our estimate of the optimal profit is <2% off the optimal profit drawn from BDM data. Our FULL de-biasing procedure has indeed delivered a rather sizeable improvement all around.

4.2. Study 2: university sweatshirt

The previous gym bag study demonstrated the promise of our de-biasing procedure for the direct single question approach. In this second study, we applied the same procedure to a different data set we collected to investigate the robustness of our de-biasing approach. Here, we report a summary of the economic analysis of Study 2 and refer the reader to the detailed information on this study in the Web Appendix G.

As with the gym bag study, the OE data series on a university sweatshirt generate statistically different optimal quantity and the DC data series different optimal price estimates. Here too, the traditional interpretations of both data series lead to wildly overestimated profits, 48.83% and 23.03% respectively. The first BASIC de-biasing procedure subtracting only $\tilde{p}-\hat{p}$ improves significantly on the original OE data, but does not beat the DC data series. The second, EPSILON, where individual-specific variation is accounted for, improves on DC significantly and the profit estimate is different from that of BDM by only 6.16%. Our FULL de-biasing procedure performs, once again, remarkably well. The estimates of the optimal price and quantity are not statistically different from BDM and the estimate of optimal profits is within 5.26% of the BDM.

4.3. Selecting the Right $cov(\theta,p)$

From both studies, we have seen that our FULL de-biasing procedure statistically improved the OE data series in a robust way. More importantly, the de-biased data series using the FULL procedure always generated much improved estimates for the optimal price, quantity, and profits, judging by the gold standard of BDM estimates. These improvements can help a firm make better pricing decisions, better production planning, and better market entry decisions.

One remaining issue from both studies, which has great importance for researchers using our de-biasing procedure, is how to select a $cov(\theta,p)$. For our studies, we have simulated a wide range, $cov(\theta,p)\epsilon[-3.08,7.08]$ in the first study and $cov(\theta,p)\epsilon[-8.20,1.80]$ in the second study. As shown in Fig. 3, our simulation results indicate that the value of $cov(\theta,p)$ that yields the profit estimate closest to the BDM is respectively 2.33 and −5.64. These numbers are, of course, not surprising to us, as we know from Eq. (4) that the theoretical best value for $cov(\theta,p)$ for the purpose of our de-biasing procedure is given by $\bar{p}-\hat{p}$. In the first study, we have $\bar{p}-\hat{p}=2.08$ and in the second study, we have $\bar{p}-\hat{p}=-3.20$. As Tables 5 and F.2 show, the difference between the profit estimates at the empirical best $cov(\theta,p)$ values versus the theoretical best values is diminishingly small. Therefore, we suggest that analysts can use a small BDM sample to pin down $\bar{p}$ and then determine $cov(\theta,p)$ using DC data.

It is important to note that one needs a far smaller sample, only a handful of data points, to come up with an estimate of $\bar{p}$ (i.e., from BDM) than to estimate the optimal price, quantity, and profits. However, once a rough estimate of $\bar{p}$ is obtained, one can proceed to conduct sensitivity analysis around the $cov(\theta,p)$ implied by $\bar{p}$. The range of the $cov(\theta,p)$ used for the sensitivity analysis can be very small as the differences between the empirical best $cov(\theta,p)$ values and the theoretical best values are diminishingly small as noted earlier. In the case where even a small BDM sample is too cumbersome to gather, our two studies confirm that a partial de-biasing procedure with $\tilde{p}-\hat{p}$ subtracted from the OE data series and accounting for individual-specific variations can still significantly improve our estimates relative to using the original OE data series.

5. Conclusion

In this paper, we take a close look at two single question approaches to gauge consumers' willingness to pay (WTP), the open-ended (OE) question and the dichotomous-choice (DC) question, which are frequently used by practitioners because the data are easy to collect. These approaches are conceptually simple, easy to implement, and unfailingly generate timely information, even in real-time, at a low cost. However, these advantages are negated if these approaches elicit a hypothetical WTP that deviates significantly from the consumers' actual WTP — as is often the case. Marketing scholars have thus far failed to find ways to reduce hypothetical bias for single question approaches.

In this paper, we make the first attempt to de-bias the direct single question approach in a rigorous way. Specifically, we systematically investigate the nature of hypothetical biases associated with two basic single question formats, the open-ended (OE), and the dichotomous-choice (DC), and look for ways to leverage data from both question formats to reduce the bias in the OE question format. We do this by specifying an individual-level bias model based on theory and show that for price generation tasks such as the OE question format, a respondent tends to inflate her stated WTP, while for price selection tasks such as the DC question format, a respondent tends to show biases anchored on presented price cues. Further, we analytically derive a de-biasing procedure for the OE question format and show empirically the promise of this procedure.

Our results show that although the single question approach suffers from various kinds of statistical biases, there is no reason to discard the approach altogether. Indeed, when the market researcher uses the single question approach in combination with two question formats (i.e., the OE and DC question), she can overcome the specific biases generally associated with these formats by using our proposed de-biasing procedure. Our analysis shows that this procedure can de-bias the OE data enough to arrive at statistically and managerially valid forecasts of consumers' WTP without resorting to any BDM mechanism. When a small sample of BDM is obtainable, which is frequently the case in marketing applications, our proposed FULL procedure performs astonishingly well. Thus, our proposed de-biasing procedures preserve the advantages of simplicity and low-cost that marketing researchers seek and value in the direct single question approach.

The de-biasing procedure we have proposed shows considerable promise for improving marketing practice and we hope that the step we have taken will inspire more interest in improving single question approaches. Future research can test our de-biasing procedure in an even broader application context and may further illuminate some more subtle bias structures inherent in single question approaches.

In our investigation, we have looked into two particular lower priced consumer goods for which reference prices can be obtained. If anything, this setting renders any measured bias conservative as we believe that for such products the bias should be relatively smaller. For higher valued goods the bias tends to be higher (Murphy et al., 2005; Schmidt & Bijmolt, 2020), and this may occur because they are less frequently purchased durable goods for which price preferences are remembered less compared to more frequently purchased non-durable goods (Estelami, Lehmann, & Holden, 2001). The bias may also be higher for more novel goods and for goods for which reference prices are harder to obtain because respondents may be more susceptible to biases induced by question formats when preferences are not well-developed yet and they are less familiar with the purchased item. Similarly, the magnitude of the bias may change along a product's life cycle. It may be higher early on and become smaller over the cycle. Finally, the bias may also vary across consumer types as it has been found that consumers who are low in purchase interest and low involvement tend to yield higher biases (Hofstetter et al., 2013). In sum, although the magnitude of the bias may vary across products and consumers, its structure should be rather consistent across contexts as it is introduced by the particular OE and DC question formats. Investigating the applicability of our de-biasing approach across various contexts may be a fruitful direction of future research. Finally, future research may compare the de-biasing approach proposed in this paper with alternative de-biasing approaches in the spirit of Miller et al. (2011), aiming at improving our understanding of when which approach will yield the most valid results.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijresmar.2020.04.006.

References

Aadland, D., & Caplan, A. J. (2003). Willingness to pay for curbside recycling with detection and mitigation of hypothetical bias. *American Journal of Agricultural Economics*, *85*, 492–502.

Anderson, J. C., Jain, D., & Chintagunta, P. K. (1992). Understanding customer value in business markets: Methods of customer value assessment. *Journal of Business to-Business Marketing*, *1*(1), 3–29.

Arrow, K., Solow, R., Portney, P. R., Leamer, E. E., Radner, R., & Schuman, H. (1993). Report of the NOAA panel on contingent valuation. *Federal Register*, *58*, 4601–4614.

Balistreri, E., McClelland, G., Poe, G., & Schulze, W. (2001). Can hypothetical questions reveal true values? A laboratory comparison of dichotomous choice and open-ended contingent values with auction values. *Environmental and Resource Economics*, *18*(3), 275–292.

Becker, G. M., DeGroot, M. H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, *9*(3), 226–232.

Bettman, J. R., Luce, M. F., & Payne, J. W. (1998). Constructive consumer choice processes. *Journal of Consumer Research*, *25*(3), 187–217.

Bishop, R. C., Boyle, K. J., Carson, R. T., David, C., Michael Hanemann, W., Kanninen, B., ... Scherer, N. (2017). Putting a value on injuries to natural assets: The BP oil spill. *Science*, *356*(6335), 253–254.

Bishop, R. C., Welsh, M. P., & Heberlein, T. A. (1992). *Some experimental evidence on the validity of contingent valuation.* Working Paper Department of Agricultural Economics, University of Wisconsin, 1–27.

Brazell, J. D., Diener, C. G., Karniouchina, E., Moore, W. L., Severin, V., & Uldry, P. -F. (2006). The no-choice option and dual response choice designs. *Marketing Letters*, *17*(4), 255–268.

Brown, T. C., Champ, P. A., Bishop, R. C., & McCollum, D. W. (1996). Which response format reveals the truth about donations to a public good? *Land Economics*, *72*(2), 152–166.

Brown, T. C., Ajzen, I., & Hrubes, D. (2003). Further tests of entreaties to avoid hypothetical bias in referendum contingent valuation. *Journal of environmental Economics and Management*, *46*(2), 353–361.

Brynjolfsson, E., Collis, A., & Eggers, F. (2019). Using massive online choice experiments to measure changes in well-being. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(15), 7250–7255.

Carson, R. T., Flores, N. E., Martin, K. M., & Wright, J. L. (1996). *Source: Land Economics72. (pp. 80–99), 80–99.*

Carson, R. T. (2000). Contingent valuation: A user's guide. *Environmental Science & Technology*, *34*(8), 1413–1418.

Carson, R. T., Groves, T., & List, J. A. (2006). *Probabilistic influence and supplemental benefits a field test of the two key assumptions behind using stated preferences.* (Working Paper).

Carson, R. T., Groves, T., & Machina, M. (2000). *Incentive and informational properties of preference questions.* Working Paper Department of Economics, University of California, San Diego.

Champ, P. A., & Bishop, R. C. (2001). Donation payment mechanisms and contingent valuation: An empirical study of hypothetical bias. *Environmental and Resource Economics*, *19*(4), 383–402.

Champ, P. A., Bishop, R. C., Brown, T. C., & McCollum, D. W. (1997). Using donation mechanisms to value nonuse benefits from public goods. *Journal of Environmental Economics and Management*, *33*(2), 151–162.

Chernev, A. (2003). Reverse pricing and online price elicitation strategies in consumer choice. *Journal of Consumer Psychology*, *13*(1/2), 51–62.

Chernev, A. (2006). Decision focus and consumer choice among assortments. *Journal of Consumer Research*, *33*(1), 50–59.

Cummings, R., Harrison, G., & Osborne, L. L. (1995). *Are realistic referenda real?* Economic Working Paper B-95-06 Division of Research, College of Business Administration, University of South Caroline.

Cummings, R. G., & Taylor, L. O. (1999). Unbiased value estimates for environmental goods: A cheap talk design for the contingent valuation method. *American Economic Review*, *89*(3), 649–665.
Dickie, M., Fisher, A., & Gerking, S. (1987). Market transactions and hypothetical demand data: A comparative study. *Journal of the American Statistical Association*, *82*(3), 69–75.
Ding, M. (2007). An incentive-aligned mechanism for conjoint analysis. *Journal of Marketing Research*, *44*(May), 214–223.
Ding, M., Grewal, R., & Liechty, J. (2005). Incentive-aligned conjoint analysis. *Journal of Marketing Research*, *42*(February), 67–82.
Dong, S., Ding, M., & Huber, J. (2010). A simple mechanism to incentive-align conjoint experiments. *International Journal of Research in Marketing*, *24*(4), 312–323.
Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* Boca Raton, FL: Chapman & Hall/CRC.
Estelami, H., Lehmann, D. R., & Holden, A. C. (2001). Macro-economic determinants of consumer price knowledge: A meta-analysis of four decades of research. *International Journal of Research in Marketing*, *18*(4), 341–355.
Farrell, J., & Rabin, M. (1996). Cheap talk. *Journal of Economic Perspectives*, *10*(3), 103–118.
Fox, J. A., Shogren, J. F., Hayes, D. J., & Kliebenstein, J. B. (1998). CVM-X: Calibrating contingent values with experimental auction markets. *American Journal of Agricultural Economics*, *80*(3), 455–465.
Gabor, A., & Granger, C. W. J. (1966). Price as an indicator of quality: Report on an enquiry. *Economica*, *33*(February), 43–70.
Green, D., Jacowitz, K. E., Kahneman, D., & McFadden, D. (1998). Referendum contingent valuation, anchoring and willingness to pay for public goods. *Resource and Energy Economics*, *20*(2), 85–116.
Green, P. E., & Krieger, A. M. (1996). Modifying cluster-based segments to enhance agreement with an exogenous response variable. *Journal of Marketing Research*, *33*(3), 351–363.
Hanson, W., & Martin, K. R. (1990). Optimal bundle pricing. *Management Science*, *36*(2), 155–174.
Harrison, G. W., & Rutström, E. E. (2008). Experimental evidence on the existence of hypothetical bias in value elicitation methods. In C. R. Plott, & V. L. Smith (Eds.), *Handbook of experimental economics results.* New York: Elsevier Press.
Hoeffler, S. (2003). Measuring preferences for really new products. *Journal of Marketing Research*, *40*(4), 406–421.
Hoehn, J. P., & Randall, A. (1987). A satisfactory benefit-cost indicator from contingent valuation. *Journal of Environmental Economics and Management*, *14*(3), 226–247.
Hoffman, E., Menkhaus, D. J., Chakravarti, D., Field, R. A., & Whipple, G. D. (1993). Using laboratory experimental auctions in marketing research: A case study of new packaging for fresh beef. *Marketing Science*, *12*(3), 318–338.
Hofler, R. A., & List, J. A. (2004). Valuation on the frontier: Calibrating actual and hypothetical statements of value. *American Journal of Agricultural Economics*, *86*(1), 213–221.
Hofstetter, R., Miller, K. M., Krohmer, H., & Zhang, Z. J. (2013). How do consumer characteristics affect the bias in measuring willingness to pay for innovative products? *Journal of Product Innovation Management*, *30*(5), 1042–1053.
Jacquemet, N., Joule, R. V., Luchini, S., & Shogren, J. F. (2013). Preference elicitation under oath. *Journal of Environmental Economics and Management*, *65*, 10–13.
Jedidi, K., & Jagpal, S. (2009). Willingness to pay: Measurement and managerial implications. In V. R. Rao (Ed.), *Handbook of pricing in marketing* (pp. 37–60). Cheltenham: Edward Elgar Publishing.
Jedidi, K., & Zhang, Z. J. (2002). Augmenting conjoint analysis to estimate consumer reservation price. *Management Science*, *48*(10), 1350–1368.
John, D. R., Scott, C. A., & Bettman, J. R. (1986). Sampling data for covariation assessment: The effect of prior beliefs on search patterns. *Journal of Consumer Research*, *13*(1), 38–47.
Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgement under uncertainty: Heuristics and biases.* New York: Cambridge University Press.
Kemp, M., & Maxwell, C. (1993). Exploring a budget context for contingent evaluation. In J. Hauseman (Ed.), *Contingent valuation: A critical assessment.* Amsterdam: North-Holland.
Krinsky, I., & Robb, A. L. (1986). On approximating the statistical properties of elasticities. *Review of Economics and Statistics*, *68*(November), 715–719.
Laghaie, A., & Otter, T. (2019). *Bridging between hypothetical and incentivized choice.* (Working Paper).
Landry, C. E., & List, J. A. (2007). Using ex-ante approaches to obtain credible signals of value in contingent markets: Evidence from the field. *American Journal of Agricultural Economics*, *89*, 420–429.
Leigh, T. W., MacKay, D. B., & Summers, J. O. (1984). Reliability and validity of conjoint analysis and self-explicated weights: A comparison. *Journal of Marketing Research*, *21*(4), 456–462.
List, J. A. (2001). Do explicit warnings eliminate the hypothetical bias in elicitation procedures? Evidence from field auctions for sportscards. *American Economic Review*, *91*(5), 1498–1507.
List, J. A., & Shogren, J. F. (1998). Calibration of the difference between actual and hypothetical valuations in a field experiment. *Journal of Economic Behavior & Organization*, *37*(2), 193–205.
List, J. A., & Shogren, J. F. (2002). Calibration of willingness-to-accept. *Journal of Environmental Economics and Management*, *43*(2), 219–233.
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., Brown, T., Lucero, B., & Peterson, G. (1997). Evaluating the validity of the dichotomous choice question format in contingent valuation. *Environmental and Resource Economics*, *10*(2), 109–123.
Loomis, J., González-Cabán, A., & Robin, G. (1996). *A contingent valuation study of the value of reducing fire hazards to old-growth forests in the Pacific Northwest.* Res. Paper PSW-RP-229 *24*.Albany, CA: Pacific Southwest Research Station, Forest Service, U.S. Department of Agriculture.
Louviere, J. J., & Woodworth, G. (1983). Design and analysis of simulated consumer choice or allocation experiments: An approach based on aggregate data. *Journal of Marketing Research*, *20*(November), 350–367.
Lusk, J. L. (2003). Effects of cheap talk on consumer willingness-to-pay for golden rice. *American Journal of Agricultural Economics*, *85*(4), 840–856.
Lusk, J. L., & Schroeder, T. C. (2004). Are choice experiments incentive compatible? A test with quality differentiated beef steaks. *American Journal of Agricultural Economics*, *86*(2), 467–482.
Miller, K. M., Hofstetter, R., Krohmer, H., & Zhang, Z. J. (2011). How should consumers' willingness to pay be measured: An empirical comparison of state of the art approaches. *Journal of Marketing Research*, *48*(1), 171–184.
Mitchell, R. C., & Carson, R. T. (2013). *Using surveys to value public goods: The contingent valuation method.* Washington DC: Resources for the Future.
Monroe, K. B. (1973). Buyers' subjective perceptions of price. *Journal of Marketing Research*, *10*(1), 70–80.
Murphy, J. J., Geoffrey Allen, P., Stevens, T. H., & Weatherhead, D. (2005). A meta-analysis of hypothetical bias in stated preference valuation. *Environmental and Resource Economics*, *30*, 313–325.
Murphy, J. J., & Stevens, T. H. (2004). Contingent valuation, hypothetical bias, and experimental economics. *Agricultural and Resource Economics Review*, *33*(2), 182–192.
O'Donnell, M., & Evers, E. R. K. (2019). Preference reversals in willingness-to-pay and choice. *Journal of Consumer Research*, *45*(6), 1315–1330.
Park, Y. -H., Ding, M., & Rao, V. R. (2008). Eliciting preference for complex products: A web-based eliciting upgrading method. *Journal of Marketing Research*, *45*(October), 562–574.
Poe, G. L., Clark, J. E., Rondeau, D., & Schulze, W. D. (2002). Provision point mechanisms and field validity tests of contingent valuation. *Environmental and Resource Economics*, *23*(1), 105–131.
Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science*, *306*(5695), 462–466.
Schmidt, J., & Bijmolt, T. H. A. (2020). Accurately measuring willingness to pay for consumer goods: A meta-analysis of the hypothetical bias. *Journal of the Academy of Marketing Science*, *48*, 499–518. https://doi.org/10.1007/s11747-019-00666-6.
Shaffer, G., & Zhang, Z. J. (1995). Competitive coupon targeting. *Marketing Science*, *14*(4), 395–416.
Shaffer, G., & Zhang, Z. J. (2000). Competitive pay to switch or pay not to switch: Third degree price discrimination in markets with switching costs. *Journal of Economics & Management Strategy*, *9*(3), 397–424.
Shapiro, B. P. (1968). The psychology of pricing. *Harvard Business Review*, *46*(7), 14–25.
Shogren, J. F. (1990). The impact of self-protection and self-insurance on individual response to risk. *Journal of Risk and Uncertainty*, *3*(2), 191–204.
Simmons, J. (2014). MTurk vs. The Lab: Either way we need big samples. Retrieved from: http://datacolada.org/18.

Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior & Human Performance*, *6*(6), 649–744.
Steiner, M., & Hendus, J. (2012). *How consumers' willingness to pay is measured in practice: An empirical analysis of common approaches' relevance.* (SSRN Working Paper).
Stevens, T. H., Tabatabaei, M., & Lass, D. (2013). Oaths and hypothetical bias. *Journal of Environmental Management*, *127*, 135–141.
Sun, N., & Yang, Z. (2014). An efficient and incentive compatible dynamic auction for multiple complements. *Journal of Political Economy*, *122*(2), 422–466.
Van Westendorp, P. (1976). NSS-Price Sensitivity Meter (PSM): A new approach to study consumer perception of prices. *Proceedings of the 29th ESOMAR Congress, Amsterdam, 139-67*.
Wang, T., Venkatesh, R., & Chatterjee, R. (2007). Reservation price as a range: An incentive-compatible measurement approach. *Journal of Marketing Research*, *44*(May), 200–213.
Weaver, R., & Prelec, D. (2013). Creating truth-telling incentives with the Bayesian truth serum. *Journal of Marketing Research*, *50*(3), 289–302.
Wertenbroch, K., & Skiera, B. (2002). Measuring consumers' willingness to pay at the point of purchase. *Journal of Marketing Research*, *39*(2), 228–241.