Journal of the Academy of Marketing Science (2020) 48:499–518
https://doi.org/10.1007/s11747-019-00666-6

REVIEW PAPER

Accurately measuring willingness to pay for consumer goods: a meta-analysis of the hypothetical bias

Jonas Schmidt[1] · Tammo H. A. Bijmolt[2]

Received: 30 August 2018 / Accepted: 21 May 2019 / Published online: 7 June 2019

Abstract

Consumers' willingness to pay (WTP) is highly relevant to managers and academics, and the various direct and indirect methods used to measure it vary in their accuracy, defined as how closely the hypothetically measured WTP (HWTP) matches consumers' real WTP (RWTP). The difference between HWTP and RWTP is the "hypothetical bias." A prevalent assumption in marketing science is that indirect methods measure WTP more accurately than do direct methods. With a meta-analysis of 77 studies reported in 47 papers and resulting in 115 effect sizes, we test that assumption by assessing the hypothetical bias. The total sample consists of 24,347 included observations for HWTP and 20,656 for RWTP. Moving beyond extant meta-analyses in marketing, we introduce an effect size metric (i.e., response ratio) and a novel analysis method (i.e., multivariate mixed linear model) to analyze the stochastically dependent effect sizes. Our findings are relevant for academic researchers and managers. First, on average, the hypothetical bias is 21%, and this study provides a reference point for the expected magnitude of the hypothetical bias. Second, the deviation primarily depends on the use of a direct or indirect method for measuring HWTP. In contrast with conventional wisdom, indirect methods actually overestimate RWTP significantly stronger than direct methods. Third, the hypothetical bias is greater for higher valued products, specialty goods (cf. other product types), and within-subject designs (cf. between-subject designs), thus a stronger downward adjustment of HWTP values is necessary to reflect consumers' RWTP.

Keywords Willingness to pay · Reservation price · Pricing · Conjoint analysis · Measurement accuracy · Hypothetical bias · Meta-analysis · Response ratio · Stochastically dependent effect sizes

Introduction

In a state-of-practice study of consumer value assessments, Anderson et al. (1992, p. 3) point out that consumers' willingness to pay (WTP) is "the cornerstone of marketing strategy" that drives important marketing decisions. First, consumers' WTP is the central input for price response models that inform optimal pricing and promotion decisions. Second, a new product's introductory price must be carefully chosen, because a poorly considered introductory price can jeopardize the investments in its development and threaten innovation failures (Ingenbleek et al. 2013). Not only do companies need to know what consumers are willing to pay early in their product development process, but WTP is also of interest to researchers in marketing and economics who seek to quantify concepts such as a product's value (Steiner et al. 2016). Obtaining accurate measures of consumers' WTP thus is essential.

Existing methods for measuring WTP can be assigned to a 2 × 2 classification (Miller et al. 2011), according to whether they measure WTP in a hypothetical or real context, with direct or indirect measurement methods (see Table 1). First, a hypothetical measure of WTP (HWTP) does not impose any financial consequences for participants' decisions. Participants just state what they would pay for a product, if

Mark Houston and John Hulland served as Special Issue Editors for this article.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11747-019-00666-6) contains supplementary material, which is available to authorized users.

✉ Jonas Schmidt
jo.schmidt@uni-muenster.de

Tammo H. A. Bijmolt
t.h.a.bijmolt@rug.nl

[1] Marketing Center Muenster, University of Muenster, Am Stadtgraben 13-15, 48143 Muenster, Germany

[2] Department of Marketing, Faculty of Economics and Business, University of Groningen, Nettelbosje 2, 9747 AE Groningen, The Netherlands

Table 1 Classification of methods for measuring WTP

	Type of measurement	
Context	Direct	Indirect
Hypothetical	• *Open questioning*	
	• *Closed-ended*	• *Conjoint analysis*
	• *Choice bracketing procedure*	
Real	• *Vickrey auction*	
	• *BDM lottery*	
	• *Random n^{th} price auction*	• *Incentive-aligned conjoint analysis*
	• *English auction*	
	• *eBay*	

given the opportunity to buy it. In contrast, participants may be required to pay their stated WTP in a real context, which provides a real measure of WTP (RWTP). This could for example be in the context of an auction, where the winner in the end actually has to buy the product. The difference between RWTP and HWTP is induced by the hypothetical context and is called "hypothetical bias." This hypothetical bias provides a measure of the hypothetical method's accuracy (Harrison and Rutström 2008). In case HWTP is measured with two different methods, the one with the lower hypothetical bias gives a more accurate estimate of participants' RWTP, increasing the estimate's validity. We conceptualize the hypothetical bias as the ratio of HWTP to RWTP. A method yielding an exemplary hypothetical bias of 1.5 shows that those participants overstate their RWTP for a product by 50% when asked hypothetically. Second, direct methods ask consumers directly for their WTP, whereas indirect methods require consumers to evaluate, compare, and choose among different product alternatives, and the price attribute is just one of several characteristics. Then, WTP can be derived from their responses.

Many researchers assume that direct methods create a stronger hypothetical bias, because they evoke greater price consciousness (Völckner 2006). In their pricing textbook, Nagle and Müller (2018) allege that direct questioning "should never be accepted as a valid methodology. The results of such studies are at best useless and are potentially highly misleading" (p. 186). Simon (2018) takes a similar line, stating, "It doesn't make sense to ask consumers directly for the utility or their WTP, as they aren't able to give a direct and precise estimate. The most important method to quantify utilities and WTP is the conjoint analysis" (p. 53). Because indirect methods represent a shopping experience, they are expected to be more accurate for measuring HWTP (Breidert et al. 2006; Leigh et al. 1984; Völckner 2006). Still, practitioners largely continue to rely on direct survey methods, which tend to be easier to implement (Anderson et al. 1992; Hofstetter et al. 2013; Steiner and Hendus 2012).

Various studies specify the accuracy of one or more direct or indirect methods by comparing HWTP with RWTP. Yet no clear summary of these findings is available,[1] and considering the discrepancy between theory and practice, "there is a lack of consensus on the 'right' way to measure […] consumer's reservation price" (Wang et al. 2007, p. 200). Therefore, with this study we seek to shed new light on the relative accuracy of alternative methods for measuring consumers' WTP, and particularly the accuracy of direct versus indirect methods. We perform a meta-analysis of existing studies that measure HWTP and RWTP for the same product or service, which reveals some empirical generalizations regarding accuracy. We also acknowledge the potential influence of other factors on the accuracy of WTP measures (Hofstetter et al. 2013; Sichtmann et al. 2011), such that we anticipate substantial heterogeneity across extant studies. With a meta-regression, we accordingly identify moderators that might explain this heterogeneity in WTP accuracy (Thompson and Sharp 1999; van Houwelingen et al. 2002). Our multivariate mixed linear model enables us to analyze the stochastically dependent effect sizes (ESs) explicitly (Gleser and Olkin 2009; Kalaian and Raudenbush 1996), which provides the most accurate way to deal with dependent ESs (van den Noortgate et al. 2013). As an effect size (ES) measure, we use the response ratio of HWTP and RWTP (Hedges et al. 1999), such that we obtain the relative deviation of HWTP. To the best of our knowledge, no previous meta-analysis in marketing has applied a mixed linear model nor a response ratio to measure ESs.

On average, the hypothetical bias is about 21%. In addition, direct methods outperform indirect methods with regard to their accuracy. The meta-regression shows that, compared with direct measurement methods, the hypothetical bias is considerably higher in indirect measures, by 10 percentage

[1] Three meta-analyses dealing with the hypothetical bias exist (Carson et al. 1996; List and Gallet 2001; Murphy et al. 2005). However, they focus on public goods and their results are of limited use for marketing. In contrast to the existing meta-analyses, we focus on private goods and include several private good specific moderators of high interest for marketers. For a more detailed discussion of the three existing meta-analyses, please refer to Web Appendix A.

points in a full model. This finding contradicts the prevailing wisdom in academic studies but supports current practices in companies. In addition to the type of measurement, value of the product, product type, and type of subject design have a significant influence on the hypothetical bias.

In the next section, we prove an overview of WTP and its different measurement options. After detailing the data collection and coding, we explicate our proposed ES measure, which informs the analysis approach we take to deal with stochastically dependent ESs. We present the results and affirm their robustness with multiple methods. Finally, we conclude by highlighting our theoretical contributions, explaining the main managerial implications, and outlining some limitations and directions for further research.

Willingness to pay

Definition and classification

We take a standard economic view of WTP (or reservation price) and define it as the maximum price a consumer is willing to pay for a given quantity of a product or a service (Wertenbroch and Skiera 2002). At that price, the consumer is indifferent to buying or not buying, because WTP reflects the product's inherent value in monetary terms. That is, the product and the money have the same value, so spending to obtain a product is the same as keeping the money.

Hypothetical versus real WTP

The first dimension in Table 1 distinguishes between hypothetical and real contexts, according to whether the measure includes a payment obligation or not. Most measures of RWTP rely on incentive-compatible methods, which ensure it is the participant's best option to reveal his or her true WTP. Several different incentive-compatible methods are available (Noussair et al. 2004) and have been used in prior empirical studies to measure RWTP. However, all methods that measure RWTP require a finished, sellable version of the product. Therefore, practitioners regularly turn to HWTP during the product development process, before the final product actually exists. In addition, measuring RWTP can be difficult and expensive, for both practitioners and researchers. Therefore, the accuracy of HWTP methods is of interest to practitioners and academics alike. Because RWTP reflects consumers' actual valuation of a product, it provides a clear benchmark for comparison with HWTP. We integrate existing empirical evidence about the accuracy of various direct and indirect methods to measure HWTP.

Direct methods to measure WTP

Direct measures usually include open questions, such as, "What is the maximum you would pay for this product?" Other methods use closed question formats (Völckner 2006) and require participants to state whether they would accept certain prices or not. Still others combine closed and open questions. The choice bracketing procedure starts with several closed questions, each of which depends on the previous answer. If consumers do not accept the last price of the last closed question, they must answer an open question about how much they would be willing to pay (Wertenbroch and Skiera 2002).

In particular, the most widely used direct measures of RWTP are the Vickrey auction (Vickrey 1961) and the Becker-DeGroot-Marschak lottery (BDM) (Becker et al. 1964). In a Vickrey auction, every participant hands in one sealed bid. The highest bidder wins the auction but pays only the price of the second highest bid; accordingly, these auctions also are called second-price sealed bid auctions. By disentangling the bid and the potential price, no bidding strategy is superior to bidding actual WTP. Different adaptions of these Vickrey auctions are available, such as the random nth price auction (Shogren et al. 2001), in which participants do not know the quantity being sold in the auction upfront. In contrast, a BDM lottery does not require participants to compete for the product. Instead, participants first state their WTP, and then a price is drawn randomly. If her or his stated WTP is equal to or more than the drawn price, a participant must buy the product for the drawn price. If the stated WTP is less than the drawn price, she or he may not buy the product. Similar to the Vickrey auction, the stated WTP does not influence the drawn price and therefore does not determine the final price. Again then, the dominant strategy is to state actual WTP.

Not all direct measures of RWTP are theoretically incentive compatible. For example, in an English auction, the price increases until only one interested buyer is left, who eventually buys the product for the highest announced bid. Every bidder has an incentive to bid up WTP (Rutström 1998), so an English auction reveals all bidders' WTP, except for the winner's, who stops bidding after the last competitor leaves. Therefore, the English auction is not theoretically incentive compatible, yet the mean RWTP obtained tend to be similar to those resulting from incentive-compatible methods (Kagel et al. 1987). Therefore, we treat studies using an English auction as direct measures of RWTP.

Finally, the online auction platform eBay can provide a direct measure of RWTP. Unlike a Vickrey auction, the auction format implemented in eBay allows participants to bid multiple times, and the auction has a fixed endpoint. Although multiple bids from one participant imply that not every bid reveals true WTP, the highest and latest bid does provide this information (Ockenfels and Roth

2006). Theoretically then, eBay auctions are not incentive compatible either (Barrot et al. 2010), but the empirical results from eBay and Vickrey auctions are highly comparable (Ariely et al. 2005; Bolton and Ockenfels 2014). Schlag (2008) gauges RWTP from eBay by exclusively using the highest bid from each participant but disregarding the winners' bid. We include this study in our meta-analysis as an example of a direct method.

Indirect methods to measure WTP

Among the variety of indirect methods to compute WTP (Lusk and Schroeder 2004), the most prominent is choice-based conjoint (CBC) analysis. Each participant chooses several times among multiple alternative products, including a "no choice" option that indicates the participant does not like any of the offered products. Each product features several product attributes, and each attribute offers various levels. To measure WTP, price must be one of the attributes. From the collected choices, it is possible to compute individual utilities for each presented attribute level and, by interpolation, each intermediate value. Ultimately, WTP can be derived according to the following relationship (Kohli and Mahajan 1991), which is the most often used approach in the studies included in the meta-analysis:

$$u_{it|-p} + u_i(p) \geq u_i^*,$$

where $u_{it\,|\,-p}$ is the utility of product t excluding the utility of the price, and $u_i(p)$ is the utility for a price level p for consumer i. In accordance with Miller et al. (2011) and Jedidi and Zhang (2002), we define u_i^* as the utility of the "no choice" option. The resulting WTP indicates the highest price p that still fulfills the relationship. In their web appendix, Miller et al. (2011) provide a numerical example.

In principle, indirect methods provide measures of HWTP, because the choices and other judgments expressed by the participants do not have any financial consequences. Efforts to measure RWTP indirectly attempt to insert a downstream mechanism that introduces a binding element (Wlömert and Eggers 2016). For example, Ding et al. (2005) propose to randomly choose one of the selected alternatives and make that choice binding. Every choice could be the binding one, so participants have an incentive to reveal their true preferences throughout the task. Ding (2007) also incorporates the idea of the BDM lottery, proposing that participants could take part in a conjoint task, from which it is possible to infer their WTP for one specific product, according to the person's choices in the conjoint task. The inferred WTP then enters the BDM lottery subsequently, so participants have an incentive to reveal their true preferences in the conjoint task.

Hypotheses

We predict that several moderators may affect the hypothetical bias. In addition, we control for several variables. The potential moderators constitute four main categories: (1) methods for measuring WTP, (2) research stimulus, (3) general research design of the study, and (4) the publication in which the study appeared. The last category only contains control variables.

Moderators: HWTP measurement

Direct methods for measuring HWTP have some theoretical drawbacks compared to indirect methods. First, asking consumers directly for their HWTP tends to prime them to focus on the price (Breidert et al. 2006), which is unlike a natural shopping experience in which consumers choose among several products that vary on multiple attributes. That is, direct methods may cause atypically high price consciousness (Völckner 2006). Indirect methods address this drawback by forcing participants to weigh the costs and benefits of different alternatives. Second, when asked directly, consumers might try to answer strategically if they suspect their answers might influence future retail prices (Jedidi and Jagpal 2009). Because indirect methods do not prompt participants to state their HWTP directly, strategic answering may be less likely. Third, direct statements of HWTP are cognitively challenging, whereas methods that mimic realistic shopping experiences require less cognitive effort (Brown et al. 1996).

Indirect methods for measuring HWTP also have some drawbacks that might influence the hypothetical bias. First, researchers using a CBC must take care to avoid a number-of-levels effect, especially in pricing studies (Eggers and Sattler 2009). To do so, they generally can test only a few different prices, which might decrease accuracy if the limitation excludes the HWTP of people with higher (lower) WTP than the highest (lowest) price shown. Second, indirect methods assume a linear relationship between price levels, through their use of linear interpolation (Jedidi and Zhang 2002).

Overall then, measuring HWTP with direct or indirect methods could evoke the hypothetical bias, and extant evidence is mixed (e.g. Miller et al. 2011), featuring arguments for the superiority of both method types. Therefore, we formulate two competing hypotheses.

H1a: Measuring HWTP with an indirect method leads to a smaller hypothetical bias compared to direct methods.

H1b: Measuring HWTP with a direct method leads to a smaller hypothetical bias compared to indirect methods.

Moderators: research stimulus

When asked for their HWTP, personal budget constraints do not exert an effect, because the consumer does not actually have to pay any money. However, when measuring RWTP, budget constraints limit the amount that participants may contribute (Brown et al. 2003). For low-priced products, this constraint should have little influence on the hypothetical bias, because the RWTP likely falls within this budget. For high-priced products though, budget constraints likely become more relevant; participants might state HWTP estimates that they could not afford in reality, thereby increasing the hypothetical bias. Thus, we hypothesize:

H2: The hypothetical bias is greater for products with a higher value.

A classic categorization of consumer goods cites convenience, shopping, and specialty goods, depending on the amount of search and price comparison effort they require (Copeland 1923). Consumers engage in more search effort when they have trouble assessing a product's utility. Hofstetter et al. (2013) in turn show that the hypothetical bias decreases as people gain means to assess a product's utility, and in a parallel finding, Sichtmann et al. (2011) show that higher product involvement reduces the hypothetical bias. That is, higher product involvement likely reduces the need for intensive search effort. Therefore, we hypothesize:

H3: The hypothetical bias is least for convenience goods, greater for shopping goods, and greatest for specialty goods.

Consumers face uncertainty about an innovative product's performance and their preferences for it (Hoeffler 2003). According to Sichtmann et al. (2011), stronger consumer preferences lower the hypothetical bias. In contrast, greater uncertainty reduces their ability to assess a product's utility, which increases the hypothetical bias (Hofstetter et al. 2013). Finally, Hofstetter et al. (2013) show that the perceived innovativeness of a product increases the hypothetical bias. Consequently,

H4: The hypothetical bias is greater for innovations compared to established products.

Moderators: research design

The research design also might influence the hypothetical bias (List and Gallet 2001; Murphy et al. 2005). In particular, the subject design of an experiment determines the results, in the sense that between-subject designs tend to be more conservative (Charness et al. 2012), whereas within-subject designs tend to result in stronger effects (Ariely et al. 2006). Fox and Tversky (1995) identify stronger effects for a within-subject versus between-subject design in the context of ambiguity aversion; Ariely et al. (2006) similarly find such stronger effects for a within-subject design for a study comparing WTP and willingness to accept. According to Frederick and Fischhoff (1998), participants in a within-subject design express greater WTP differences for small versus large quantities of a product than do those in a between-subject design. Therefore,

H5: The hypothetical bias is greater for within-subject designs compared with between-subject designs.

Another source of uncertainty pertains to product performance, and it increases when the consumer can only review images (e.g., online) rather than inspect the product itself physically (Dimoka et al. 2012). Consequently, many consumers test products in a store to reduce their uncertainty before buying them online (showrooming) (Gensler et al. 2017). Similarly, consumers' uncertainty might be reduced in a WTP experiment by giving them an opportunity to inspect and test the product before bidding. Bushong et al. (2010) show that participants state a higher RWTP when real products, rather than images, have been displayed. As Hofstetter et al. (2013) note, greater uncertainty increases the hypothetical bias. We hypothesize:

H6: Giving participants the opportunity to test a product before bidding reduces the hypothetical bias.

Finally, researchers often motivate participation in an experiment by paying some remuneration or providing an initial balance to bid in an auction. Equipping participants with money might change their RWTP, because they gain an additional budget. They even might consider this additional budget like a coupon, which they add to their original RWTP. Consumers in general overstate their WTP in hypothetical contexts, so providing a participation fee could decrease the hypothetical bias. Yet Hensher (2010) criticizes the use of participation fees, noting that they can bias participants' RWTP.

H7: Providing participants (a) a participation fee or (b) an initial balance decreases the hypothetical bias.

Collection and coding of studies

Collection of studies

With our meta-analysis, we aim to generalize empirical findings about the relative accuracy of HWTP measures, so we

conducted a search for studies that report ESs of these measures. We used three inclusion criteria. First, the study had to measure consumers' HWTP and RWTP for the same product or service, so that we could determine the hypothetical bias. Second, the research stimulus had to be private goods or services. Third, we included only studies that reported the mean and standard deviation (or values that allow us to compute it) of HWTP and RWTP or for which the authors provided these values at our request.

To identify relevant studies, we applied a keyword search in different established online databases (e.g., Science Direct, EBSCO) and Google Scholar across all research disciplines and years. The keywords included "willingness-to-pay," "reservation price," "hypothetical bias," and "conjoint analysis." We also conducted a manual search among leading marketing and economics journals. To reduce the risk of a publication bias, we extended our search to the Social Science Research Network, Research Papers in Economics, and the Researchgate network, and we checked for relevant dissertations whose results had not been published in journals. Moreover, we conducted a cross-reference search to find other studies. We contacted authors of studies that did not report all relevant values and asked them for any further relevant studies they might have conducted. Ultimately, we identified 77 studies reported in 47 articles, accounting for 117 ESs and total sample sizes of 24,441 for HWTP and 20,766 for RWTP.

Coding

As mentioned previously and as indicated by Table 2, we classify the moderators into four categories: (1) methods for measuring WTP, (2) research stimulus, (3) general research design of the study, and (4) the publication in which the study appears. In the first category, the main moderator of interest is the *type of measurement HWTP*, that is, the direct versus indirect measurement of HWTP. Two other moderators deal with RWTP measurement. *Type of measurement RWTP* similarly distinguishes between direct and indirect measures, whereas *incentive compatible* reflects the incentive compatibility (or not) of the method.

The second category of moderators, dealing with the research stimulus, includes *value*, or the mean RWTP for the corresponding product. The experiments in our meta-analysis span different countries and years, so we converted all values into U.S. dollars using the corresponding exchange rates. The variable *variance ES* captures participants' uncertainty and heterogeneity when evaluating a product. With regard to the products, we checked whether they were described as new to the consumer or innovations, which enabled us to code the *innovation* moderator. The moderator *product/service* distinguishes products and services. Finally, the *product type* moderator requires more subjective judgment. Two independent coders, unaware of the research project, coded product type by using Copeland's (1923) classification of consumer goods according to the search and price comparison effort they require, as convenience goods, shopping goods, or specialty goods. We use an ordinal scale for *product type* and therefore assessed interrater reliability with a two-way mixed, consistency-based, average-measure intraclass correlation coefficient (ICC) (Hallgren 2012). The resulting ICC of 0.82 is rated as excellent (Cicchetti 1994); the two independent coders agreed on most stimuli. The lack of any substantial measurement error indicates no notable influence on the statistical power of the subsequent analyses (Hallgren 2012). Any inconsistent codes were resolved through discussion between the two coders. We include *product type* in the analyses with two dummy variables for shopping and specialty goods, and convenience goods are captured by the intercept.

In the third category, we consider moderators that deal with the research design. The *type of experiment HWTP* and *type of experiment RWTP* capture whether the studies measure HWTP and RWTP in field or lab experiments, respectively. Experiments conducted during a lecture or class are designated lab experiments. *Offline/online HWTP* and *offline/online RWTP* indicate whether the experiment is conducted online or offline; the *type of subject design* reveals if researchers used a between- or within-subject design. The moderator *opportunity to test* indicates whether participants could inspect the product in more detail before bidding. *Participation fee* and *initial balance* capture whether participants received money for showing up or for spending in the auction, respectively. We identify a *student sample* when the sample consists of exclusively students; mixed samples are coded as not a student sample. Methods for measuring RWTP often are not self-explanatory, so researchers introduce them to participants, using various types of instruction. We focused on whether incentive compatibility concepts or the dominant bidding strategy were explained, using a moderator *introduction of method for RWTP* with four values. It equals "none" if the method was not introduced, "explanation" if the method and its characteristics were explained, "training" if mock auctions or questions designed to understand the mechanism occurred before the focal auction took place or questions were asked, and "not mentioned" if the study does not indicate whether the method was introduced. With this nominal scale, we include this moderator by using three dummy variables for explanation, training, and not mentioned, while the none category is captured by the intercept. Finally, we include *region*. Almost all the studies were conducted in North America or Europe; we distinguish North America from "other countries (mostly Europe)."

The fourth category of moderators contains publication characteristics. We checked whether a study underwent a peer review process (*peer reviewed*), reflected a marketing or economics research domain (*discipline*), how many citations it

Table 2 Moderators

Category	Moderator	Values	Variables	Description
WTP measurement	Type of measurement HWTP	Direct Indirect	Dummy variable (indirect = 1)	Whether HWTP is measured directly or indirectly.
	Type of measurement RWTP	Direct Indirect	Dummy variable (indirect = 1)	Whether RWTP is measured directly or indirectly.
	Incentive compatible	No Yes	Dummy variable (yes = 1)	Whether the method for measuring RWTP is incentive compatible.
Research stimulus	Value		Metric variable	The mean RWTP converted into US dollars.
	Product type	Convenience goods Shopping goods Specialty goods	Two dummy variables for shopping and specialty goods; convenience goods are captured by the intercept	Classification of respective stimulus based on an Copeland (1923).
	Innovation	No Yes	Dummy variable (yes = 1)	Whether the stimulus is an innovation.
	Product/service	Product Service	Dummy variable (service = 1)	Whether the stimulus is a product or a service.
	Variance ES		Metric variable	The variance of the ES.
Research design	Type of subject design	Between Within	Dummy variable (within = 1)	Whether it is a between or a within subject design.
	Opportunity to test	No Yes	Dummy variable (yes = 1)	Whether participants had the chance to test the product before bidding.
	Participation fee	No Yes	Dummy variable (yes = 1)	Whether participants received a participation fee.
	Initial balance	No Yes	Dummy variable (yes = 1)	Whether participants received an initial balance for the auction.
	Type of experiment HWTP	Field Lab	Dummy variable (lab = 1)	Whether HWTP is measured in a field or a lab experiment.
	Type of experiment RWTP	Field Lab	Dummy variable (lab = 1)	Whether RWTP is measured in a field or a lab experiment.
	Offline/online HWTP	Offline Online	Dummy variable (online = 1)	Whether HWTP is measured offline or online.
	Offline/online RWTP	Offline Online	Dummy variable (online = 1)	Whether RWTP is measured offline or online.
	Student sample	No Yes	Dummy variable (yes = 1)	Whether the sample consists of students only.
	Introduction of method for RWTP	None Explanation Training Not mentioned	Three dummy variables for explanation, training, and not mentioned; None is captured by the intercept	How the method for measuring RWTP was introduced.
	Region	Other Countries (mostly Europe) North America	Dummy variable (North America = 1)	Region where the experiment was conducted.

Table 2 (continued)

Category	Moderator	Values	Variables	Description
Publication characteristics	*Peer reviewed*	No Yes	Dummy variable (yes = 1)	Whether the study was peer reviewed.
	Discipline	Economics Marketing	Dummy variable (marketing = 1)	Corresponding research discipline
	Citations		Metric variable	Number of citations in Google Scholar
	Year		Metric variable	Year the study was published

Moderators in italics are control variables

had on Google Scholar (*citations*), and in which year it was published (*year*).

Methodology

Effect size

To determine the hypothetical bias induced by different methods, we need an ES that represents the difference between obtained values for HWTP and RWTP. When the differences stem from a comparison of a treatment and a control group, standardized mean differences (SMD) are appropriate measures (e.g. Abraham and Hamilton 2018; Scheibehenne et al. 2010). Specifically, to compute SMD, researchers divide the difference in the means of the treatment and the control group by the standard deviation, which helps to control for differences in the scales of the dependent variables in the experiments. Accordingly, it applies to studies that measure the same outcome on different scales (Borenstein et al. 2009, p. 25). In contrast, the ESs in our meta-analysis rely on the same scale; they differ in their position on the scale, because the products evoke different WTP values. In this case, the standard deviation depends on not only the scale range but also many other relevant factors, so the standard deviation should not be used to standardize the outcomes. In addition, as studies may have used alternate experimental designs, different standard deviations could be used across studies, leading to standardized mean differences that are not directly comparable (Morris and DeShon 2002). Rather than the SMD, we therefore use a response ratio to assess ES, because it depends on the group means only.

Specifically, the response ratio is the mean outcome in an experimental group divided by that in a corresponding control group, such that it quantifies the percentage of variation between the experimental and control groups (Hedges et al. 1999). Unlike SMD, the response ratio applies when the outcome is measured on a ratio scale with a natural zero point, such as length or money (Borenstein et al. 2009). Accordingly, the response ratio often assesses ES in meta-analyses in ecology domains (Koricheva and Gurevitch 2014), for which many outcomes can be measured on ratio scales. To the best of our knowledge though, the response ratio has not been adopted in meta-analyses in marketing yet. However, it is common practice to specify a multiplicative, instead of a linear, model when assessing the effects of marketing instruments on product sales or other outcomes (Leeflang et al. 2015). Hence, it would be a natural option to use an effect measure representing proportionate changes, instead of additive changes, when deriving empirical generalizations on marketing subjects like response effects to mailing campaigns. For our effort, we define the response ratio as

$$response\ ratio = \frac{\mu_{HWTP}}{\mu_{RWTP}},$$

where μ_{HWTP} and μ_{RWTP} are the means of a study's corresponding HWTP and RWTP values.

For three reasons, we run statistical analyses using the natural logarithm of the response ratio as the dependent variable. First, the use of the natural logarithm linearizes the metric, so deviations in the numerator and denominator have the same impact (Hedges et al. 1999). Second, the parameters (β) for the moderating effects in the meta-regression are easy to interpret, as a multiplication factor, by taking the exponent of the estimate ($Exp(\beta)$). Most moderators are dummy variables, and a change of the corresponding dummy value results in a change of $(Exp(\beta) - 1) * 100\%$ in the hypothetical bias. However, this point should not be taken to mean that the difference of the hypothetical bias between two conditions of a moderator is $Exp(\beta) - 1$ percentage points, because that value depends on the values of other moderators. Third, the distribution of the natural logarithm of response ratios is approximately normally distributed (Hedges et al. 1999). Consequently, we define ES as:

$$ES = \ln\left(\frac{\mu_{HWTP}}{\mu_{RWTP}}\right).$$

Modeling stochastically dependent effect sizes explicitly

Most meta-analyses assume the statistical independence of observed ESs, but this assumption only applies to limited cases; often, ESs are stochastically dependent. Two main types of dependencies arise between studies and ESs. First, studies can measure and compare several treatments or variants of a type of treatment against a common control. In our context, for example, a study might measure HWTP with different methods and compare the results to the same RWTP, leading to multiple ESs that correlate because they share the same RWTP. Treating them as independent would erroneously add RWTP to the analysis twice. This type of study is called a multiple-treatment study (Gleser and Olkin 2009). Second, studies can produce several dependent ESs by obtaining more than one measure from each participant. For example, a study might measure HWTP and RWTP for several products from the same sample. The resulting ESs correlate, because they are based on a common subject. This scenario represents a multiple-endpoint study (Gleser and Olkin 2009).

There are different approaches for dealing with stochastically dependent ESs, such as ignoring or avoiding dependence, or else modeling dependence stochastically or explicitly (Bijmolt and Pieters 2001; van den Noortgate et al. 2013). In marketing research, it is still common, and also suggested to avoid dependent ESs (Grewal et al. 2017). However, nested data structures and the associated dependent ESs are prominent in marketing research, so Bijmolt and Pieters (2001) suggest using a three-level model to account for dependency, by adding error terms on all levels. In turn, marketing researchers started to model dependence stochastically by applying multi-level regression models (e.g. Abraham and Hamilton 2018; Arts et al. 2011; Babić Rosario et al. 2016; Bijmolt et al. 2005; Edeling and Fischer 2016; Edeling and Himme 2018). However, when additional information about correlations among the ESs are available, it is most accurate to model dependence explicitly by incorporating the dependencies in the covariance matrix at the within-study level (Gleser and Olkin 2009). In contrast to modeling dependence stochastically, the covariances are not estimated but rather are calculated on the basis of the provided information. To the best of our knowledge, this approach has not been applied by meta-analyses in marketing previously.

To model stochastic dependence among ESs explicitly, we follow Kalaian and Raudenbush (1996) and use a multivariate mixed linear model with two levels: a within-studies level and a between-studies level. On the former, we estimate a complete vector of the corresponding K true ESs, $\alpha_i = (\alpha_{1i}, \ldots, \alpha_{Ki})^T$, for each study i. However, not every study examines all possible K ESs, so the vector of ES estimates for study i, $ES_i = (ES_{1i}, \ldots, ES_{L_i i})^T$, contains L_i of the total possible K ESs, and by definition, $L_i \leq K$. That is, K equals the maximum number of dependent ESs in one study (i.e., six in our sample), and every vector ES_i contains between one and six estimates. The first-level model regresses α_{ki} on ES_i with an indicator variable Z_{lki}, which equals 1 if ES_{li} estimates α_{ki} and 0 otherwise, according to the following linear model:

$$ES_{li} = \sum_{k=1}^{K} \alpha_{ki} Z_{lki} + e_{li},$$

or in matrix notation,

$$ES_i = Z_i \alpha_i + e_i.$$

The first-level errors e_i are assumed to be multivariate normal in their distribution, such that $e_i \sim N(0, V_i)$, where V_i is a $K_i \times K_i$ covariance matrix for study i, or the multivariate extension of the V-known model for the meta-regression. The elements of V_i must be calculated according to the chosen ES measure (see Web Appendix B; Gleser and Olkin 2009; Lajeunesse 2011). In turn, they form the basis for modeling the dependent ESs appropriately. The vector α_i of a study's true ES is estimated by weighted least squares, and each observation is weighted by the inverse of the corresponding covariance matrix (Gleser and Olkin 2009).

The linear model for the second stage is

$$\alpha_{ki} = \beta_{k0} + \sum_{m=1}^{M_k} \beta_{km} X_{mi} + u_{ki},$$

or in matrix notation

$$\alpha_i = X_i\beta + u_i,$$

where the K ESs become the dependent variable. The residuals u_{ki} are assumed to be K-variate normal with zero average and a covariance matrix τ. Then X_i reflects the moderator variables. By combining both levels, the resulting model is

$$ES_i = Z_iX_i\beta + Z_iu_i + e_i.$$

Estimates for τ are based on restricted maximum likelihood. The analysis uses the metafor package for meta-analyses in R (Viechtbauer 2010).

Data screening and descriptive statistics

One of the criticisms of meta-analyses is the risk of publication bias, such that all the included ESs would reflect the non-random sampling procedure. Including unpublished studies can address this concern; in our sample, 22 of 117 ESs come from unpublished studies, for an unpublished work proportion of 19%, which favorably compares with other meta-analyses pertaining to pricing, such as 10% in Tully and Winer (2014), 9% in Bijmolt et al. (2005), or 16% in Abraham and Hamilton (2018). The funnel plot for the sample, as depicted in Fig. 1, is symmetric, which indicates the absence of a publication bias. Finally, as the competing H1a and H1b indicate, we do not expect a strong selection mechanism in research or publication processes that would favor significant or high (or low) ESs. Thus, we do not consider publication bias a serious concern for our study.

To detect outliers in the data, we checked for extreme ESs using the boxplot (see Web Appendix D, Figure WA2). We are especially interested in the moderator *type of measurement HWTP*, so we computed separate boxplots for the direct and indirect measures of HWTP and thereby identified one observation for each measurement type (indirect Kimenju et al. 2005; direct Neill et al. 1994) for which the ESs (0.9079; 0.9582) exceeded the upper whisker, defined as the 75% quantile plus 1.5 times the box length. Kimenju et al. (2005) report HWTP ($11.68) values from an indirect method that overestimate RWTP ($94.48) by a factor of eight; we excluded it from our analyses. Neill et al. (1994) report HWTP ($109) that overestimates RWTP ($12) by a factor of nine when excluding outliers, and it is another outlier in our database. Thus, we excluded two of 117 observations, or less than 5% of the full sample, which is a reasonable range (Cohen et al. 2003, p. 397).

The remaining 115 ESs represent 77 studies reported by 47 different articles, with a total sample size of 24,347 for HWTP and 20,656 for RWTP. Sixteen out of these 115 ESs indicate

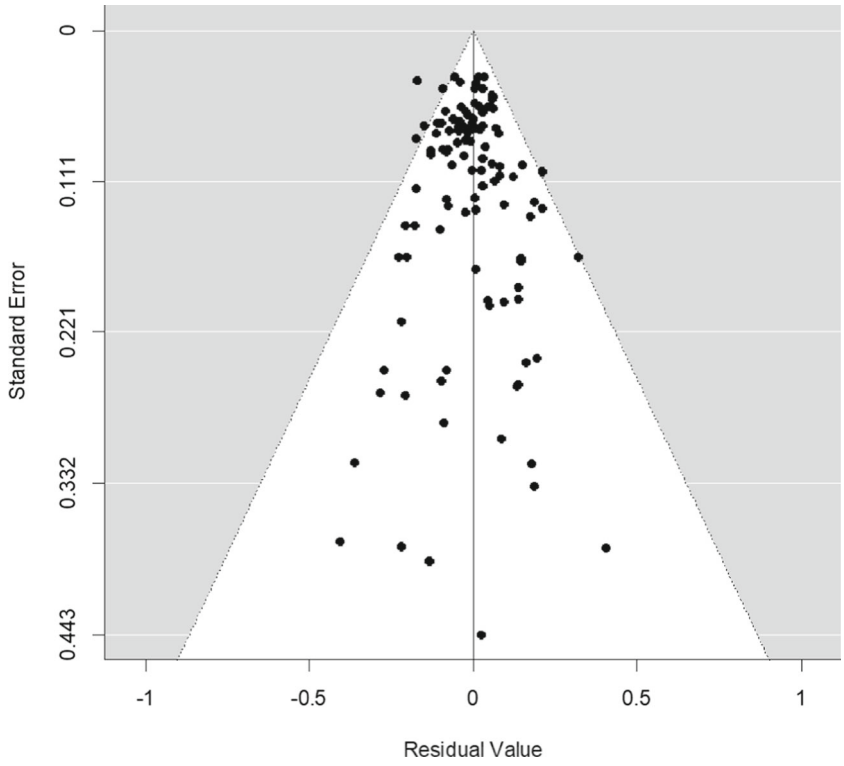


Fig. 1 Funnel plot

Notes: Six ESs with a very high standard error are not included here, to improve readability. A funnel plot with all ESs in Web Appendix C confirms the lack of a publication bias.

Table 3 Descriptive statistics

	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N
Type of measurement HWTP	Direct			Indirect								
	0.1818	0.1709	85	0.2280	0.2048	30						
Type of measurement RWTP	Direct			Indirect								
	0.1869	0.1776	106	0.2758	0.2055	9						
Incentive compatible	No			Yes								
	0.1294	0.1709	24	0.2109	0.1801	91						
Product type	Convenience			Shopping			Specialty					
	0.1954	0.1852	38	0.1339	0.1554	48	0.2911	0.1758	29			
Innovation	No			Yes								
	0.1760	0.1807	76	0.2287	0.1773	39						
Product/service	Product			Service								
	0.2482	0.1797	80	0.0696	0.1840	35						
Type of subject design	Between			Within								
	0.1800	0.1740	42	0.1800	0.1740	73						
Opportunity to test	No			Yes								
	0.1626	0.1746	75	0.2524	0.1789	40						
Participation fee	No			Yes								
	0.1400	0.1731	106	0.2747	0.1617	9						
Initial balance	No			Yes								
	0.1774	0.1662	69	0.3879	0.2365	46						
Type of experiment HWTP	Field			Lab								
	0.2716	0.1663	42	0.1491	0.1741	73						
Type of experiment RWTP	Field			Lab								
	0.2743	0.1663	39	0.1526	0.1741	76						
Offline/online HWTP	Offline			Online								
	0.1888	0.1893	87	0.2096	0.1521	28						
Offline/online RWTP	Offline			Online								
	0.1880	0.1857	91	0.2159	0.1612	24						
Student sample	No			Yes								
	0.2635	0.1571	57	0.1254	0.1769	58						
Introduction of method for RWTP	None			Explanation			Training			Not mentioned		
	0.1689	0.1670	17	0.1657	0.1863	65	0.3464	0.2096	12	0.2201	0.1144	22
Region	Other countries (mostly Europe)			North America								
	0.2678	0.1773	32	0.1653	0.1746	83						
Peer reviewed	No			Yes								

Table 3 (continued)

	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N
	0.1843	0.1938	21	0.1960	0.1785	94						
Discipline	Economics			Marketing								
	0.1194	0.1435	65	0.2907	0.1789	50						

Moderators in italics are control variables

an underestimation of RWTP, resulting from direct (12) and indirect (4) methods. Table 3 contains an overview of the moderators' descriptive statistics. *Type of measurement HWTP* reveals some mean differences between direct (0.1818) and indirect (0.2280) measures, which represents model-free support for H1b. The descriptive statistics of *product type* suggest a higher mean ES for specialty goods (0.2911) than convenience (0.1954) or shopping (0.1399) goods, in accordance with H3. With regard to *innovation*, we find a higher ES mean for innovative (0.2287) compared with non-innovative (0.1760) products, as we predicted in H4. Model-free evidence gathered from the moderators that reflect the research design also supports H5, in that the mean for between-subject designs is lower (0.1800) than that for within-subject designs (0.2798). The descriptive statistics cannot confirm H6 though, because giving participants an opportunity to test a product before stating their WTP increases the ES (0.2525) relatively to no such opportunity (0.1626). We also do not find support for H7 in the model-free evidence, because studies with an *initial balance* and *participation fee* report higher ESs than those without.

After detecting outliers and before conducting the meta-regressions, we checked for multicollinearity by calculating the generalized variance inflation factor $GVIF^{1/(2*df)}$, which is used when there are dummy regressors from categorical variables; it is comparable to the square root of the variance inflation factor $\left(\sqrt{VIF}\right)$ for 1 degree of freedom (df = 1) (Fox and Monette 1992). In an iterative procedure, we excluded the moderator with the highest $GVIF^{1/(2*df)}$ and reestimated the model repeatedly, until all moderators had a $GVIF^{1/(2*df)} < 2$. This cut-off value of 2 has been applied in other disciplines (Pebsworth et al. 2012; Vega et al. 2010) and is comparable to a VIF cut-off value of 4, within the range of suggested values (i.e., 3–5; Hair Jr et al., 2019, p. 316). Accordingly, we excluded moderators—all control variables that do not appear in any hypotheses—in the following order: *type of experiment HWTP* ($GVIF^{1/(2*df)} = 3.4723$), *offline/online RWTP* ($GVIF^{1/(2*df)} = 3.2504$), *discipline* ($GVIF^{1/(2*df)} = 2.2.4791$), *product/service* ($GVIF^{1/(2*df)} = 2.2.3290$), and *peer reviewed* ($GVIF^{1/(2*df)} = 2.0419$).

Results

To address our research questions about the accuracy of WTP measurement methods and the moderators of this performance, we performed several meta-regressions in which we varied the moderating effects included in the models. First, we ran an analysis without any moderators. Second, we ran a meta-regression with all the moderators that met the multicollinearity criteria. Third, we conducted a stepwise analysis, dropping the non-significant moderators one by one.

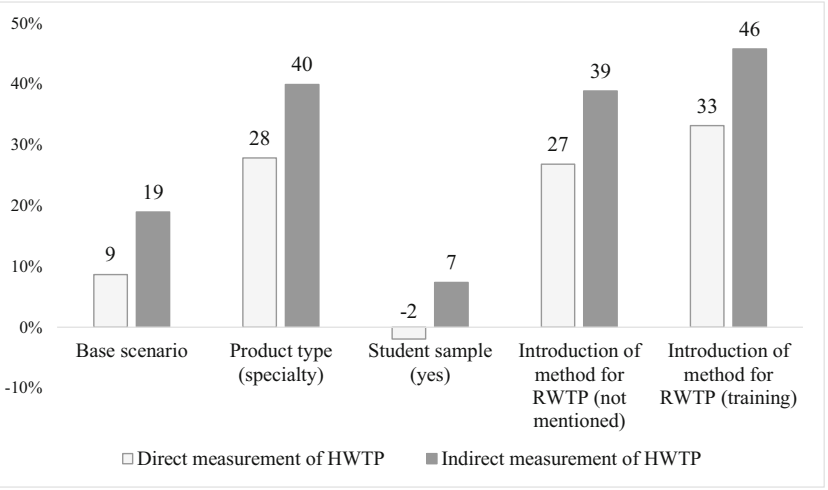
Table 4 Results of full and reduced models

	Full model					Reduced model				
	Estimate	EXP (Estimate)	Std. Err.	p Value	Significance	Estimate	EXP (Estimate)	Std. Err.	p Value	Significance
Intercept	−2.7030	0.0670	9.4731	0.7754		0.0831	1.0867	0.0500	0.0965	*
Type of measurement HWTP (indirect)	0.1027	1.1082	0.0404	0.0110	**	0.0905	1.0947	0.0382	0.0177	**
Type of measurement RWTP (indirect)	−0.0132	0.9869	0.0587	0.8216						
Incentive compatible (yes)	0.0488	1.0500	0.0574	0.3951						
Value	0.0002	1.0002	0.0001	0.0656	*					
Product type (shopping)	0.0353	1.0359	0.0445	0.4274		0.0028	1.0028	0.0371	0.9388	
Product type (specialty)	0.1615	1.1753	0.0476	0.0007	***	0.1624	1.1763	0.0393	<.0001	***
Innovation (yes)	−0.0004	0.9996	0.0505	0.9944						
Variance ES	0.1752	1.1915	0.2527	0.4883						
Type of subject design (within)	0.0878	1.0918	0.0439	0.0455	**					
Opportunity to test (yes)	0.0139	1.0140	0.0468	0.7658						
Participation fee (yes)	0.0522	1.0536	0.0489	0.2858						
Initial balance (yes)	0.0978	1.1027	0.0746	0.1896						
Type of experiment RWTP (lab)	−0.0050	0.9950	0.0471	0.9156						
Offline/online HWTP (offline)	0.0904	1.0946	0.0553	0.1019						
Student sample (yes)	−0.1134	0.8928	0.0446	0.0110	**	−0.1026	0.9025	0.0344	0.0021	***
Introduction of method for RWTP (explanation)	0.0497	1.0510	0.0579	0.3908		0.0671	1.0694	0.0420	0.1095	
Introduction of method for RWTP (training)	0.1846	1.2027	0.0762	0.0154	**	0.2032	1.2253	0.0604	0.0008	***
Introduction of method for RWTP (not mentioned)	0.1299	1.1387	0.0784	0.0974	*	0.1546	1.1672	0.0524	0.0032	***
Region (North America)	−0.0765	0.9264	0.0467	0.1013						
Citations	0.0001	1.0001	0.0001	0.3300						
Year	0.0013	1.0013	0.0047	0.7809						
τ^2	0.0031					0.0047				
R^2	0.7416					0.6083				
AICc	45.6093					−23.4892				

*Significance codes: *** p < 0.01; ** p < 0.05; * p < 0.1*

Moderators in italics are control variables

Fig. 2 Overestimation of RWTP



Notes: The base scenario is as follows: product type = convenience good,introduction of method for RWTP = explanation,student sample = no.

The first model, including only the intercept, results in an estimate (β) of 0.1889 with a standard error (SE) of 0.0183 and a p value < .0001. The estimate corresponds to an average hypothetical bias of 20.79% ($Exp(0.1889) = 1.2079$), meaning that on average, HWTP overestimates RWTP by almost 21%.

The analysis with all the moderators that met the multicollinearity threshold produces the estimation results in Table 4. The *type of measurement HWTP* has a significant, positive effect ($\beta = 0.1027$, $Exp(\beta) = 1.1082$, $SE = 0.0404$, $p = 0.0110$), indicating that indirect measures overestimate RWTP more than direct measures do. We reject H1a and confirm H1b. In particular, the ratio of HWTP to RWTP should be multiplied by 1.1082, resulting in an overestimation by indirect methods of an additional 10.82%. *Value* has a significant, positive effect at the 10% level ($\beta = 0.0002$, $Exp(\beta) = 1.0002$, $SE = 0.0001$, $p = 0.0656$), in weak support of H2. The percentage overestimation of RWTP by HWTP increases slightly, by an additional 0.02%, with each additional U.S. dollar increase in value. For H3, we find no significant difference in the hypothetical bias between convenience and shopping goods, yet specialty goods evoke a significantly higher hypothetical bias than convenience goods ($\beta = 0.1615$, $Exp(\beta) = 1.1753$, $SE = 0.0476$, $p < .0001$). This finding implies that the hypothetical bias is greater for products that demand extraordinary search effort, as we predicted in H3. We do not find support for H4, because *innovation* does not influence the hypothetical bias significantly ($\beta = -0.0004$, $Exp(\beta) = 0.9996$, $SE = 0.0505$, $p = 0.9944$).

For moderators from the research design category, we confirm the support we previously identified for H5. Measuring HWTP and RWTP using a within-subject design results in a greater hypothetical bias than does a between-subject design ($\beta = 0.0878$, $Exp(\beta) = 1.0918$, $SE = 0.0439$, $p = 0.0455$), such that the hypothetical bias increases by an additional 9.18 percentage points in this case. We do not find support for H6, H7a, or H7b though, because *opportunity to test* ($\beta = 0.0139$, $Exp(\beta) = 1.0140$, $SE = 0.0468$, $p = 0.7658$), *participation fee* ($\beta = 0.0522$, $Exp(\beta) = 1.0536$, $SE = 0.0489$, $p = 0.2858$), and *initial balance* ($\beta = 0.0978$, $Exp(\beta) = 1.1027$, $SE = 0.0746$, $p = 0.1896$) do not show significant effects.

Of the control variables, only *student sample* ($\beta = -0.1134$, $Exp(\beta) = 0.8928$, $SE = 0.0446$, $p = 0.0110$) and *introduction of method for RWTP (training)* ($\beta = 0.1846$, $Exp(\beta) = 1.2027$, $SE = 0.0762$, $p = 0.0154$) exert significant effects in the full model. If a study only includes students, the hypothetical bias gets smaller by 11%; conducting mock auctions before measuring RWTP increases the hypothetical bias by 20%.

Finally, we ran analyses in which we iteratively excluded moderators until all remaining moderators were significant at the 5% level. We excluded the moderator with the highest p value from the full model, reran the analysis, and repeated this procedure until we had only significant moderators left. We treated the dummy variables from the nominal/ordinal moderators *product type* and *introduction of method for RWTP* as belonging together, and we considered these moderators as significant when one of the corresponding dummy variables showed a significant effect. The exclusion order was as follows: *innovation*, *type of experiment RWTP*, *type of measurement RWTP*, *opportunity to test*, *year*, *variance ES*, *incentive compatible*, *initial balance*, *citations*, *participation fee*, *region*, *value*, *type of subject design*, and *offline/online HWTP*. The results in Table 4 reconfirm the support for H1b, because the *type of measurement HWTP* has a positive, significant effect ($\beta = 0.0905$, $Exp(\beta) = 1.0947$, $SE = 0.0382$, $p = 0.0177$), resulting in a multiplication factor of 1.0947. The overestimation of RWTP increases considerably for measures of WTP for specialty goods ($\beta = 0.1624$, $Exp(\beta) = 1.1763$, $SE = 0.0393$, $p < .0001$), in support of H3. Yet we do not find support for any other hypotheses in the reduced model.

Regarding the control variables, *student sample* ($\beta = -0.1026$, $Exp(\beta) = 0.9025$, $SE = 0.0344$, $p = 0.0021$) again has a significant effect, and *introduction of method for RWTP* affects the hypothetical bias significantly. In this case, the hypothetical bias increases when the article does not mention any introduction of the method for measuring RWTP to participants ($\beta = 0.1546$, $Exp(\beta) = 1.1672$, $SE = 0.0524$, $p = 0.0032$) and when the method involves mock auctions ($\beta = 0.2032$, $Exp(\beta) = 1.2253$, $SE = 0.0604$, $p = 0.0008$).

For ease of interpretation, we depict the hypothetical bias for different scenarios in Fig. 2. The reduced model provides a better model fit, according to the corrected Akaike information criterion (AICc) ($AICc_{full\ model} = 45.61$, $AICc_{reduced\ model} = -23.49$), so we use it as the basis for the simulation. The base scenario depicted in Fig. 2 measures WTP for convenience goods, explains the method for measuring RWTP to participants, and does not include solely students. The other scenarios are adaptions of the base scenario, where one of the three aforementioned characteristics is changed. In the base scenario, we predict that direct measurement overestimates RWTP by 9%, and indirect measurement overestimates it by 19%, so the difference is 10 percentage points. In contrast, for specialty goods, the overestimation increases to 28% for direct and to 40% for indirect measures. When using a pure student sample instead of a mixed sample, the predictions are relatively accurate. Here, direct measurement even underestimates RWTP by 2%, while indirect measurement yields an overestimation of 7%. With respect to how the method for measuring RWTP is introduced to the participants, not mentioning it in a paper, as well as training the method beforehand increase the hypothetical bias. While the first option is hardly interpretable, running mock tasks increases the bias to 33% in case of direct and to 46% in case of indirect methods used for measuring HWTP.

Robustness checks

We ran several additional analyses to check the robustness of the results, which we summarize in Table WA2 in Web Appendix F. To start, we analyzed Model 1 in Table WA2 by applying a cut-off value of $GVIF^{1/(2*df)} < \sqrt{10}$, comparable to the often used cut-off value of 10 for the VIF. In this case, we did not need to exclude any moderator, but the results do not deviate in their signs or significance levels relatively to the main results. *Type of measurement HWTP* still has a significant effect (5% level) on the hypothetical bias. In addition, *value*, *product type (specialty)*, and *type of subject design* exert significant influences. Among the control variables, *introduction of method for RWTP (training)*, *introduction of method for RWTP (not mentioned)*, *region*, and *peer reviewed* have significant effects (5% level). The moderators excluded from the main models due to multicollinearity (*product/service*, *type of experiment HWTP*, *offline/online RWTP*, and *discipline*) do not show significant influences.

Next, we estimated two models with all ESs, including the two outliers, but varied varied the number of included moderators (Models 2 and 3 in Table WA2). The results remain similar to our main findings. Perhaps most important, the *type of measurement HWTP* has a significant effect on the hypothetical bias, comparable in size to the effect in the main model.

In addition, instead of the multivariate mixed linear model, we used a random-effects, three-level model, such that the ES measures nested within studies with a V-known model at the lowest level (Bijmolt and Pieters 2001; van den Noortgate et al. 2013), which can account for dependence between observations. We estimated the two main models and the three robustness check models with this random-effects three-level model (Models 4–8 in Table WA2). Again, the results do not change substantially, except for *value*, which becomes significant at the 5% level.

Finally, we tested for possible interaction effects. That is, we took all significant moderators from the full model and tested, for each significant moderator, all possible interactions. The limited number of observations prevented us from simultaneously including all interactions in one model. Therefore, we first estimated separate models for each of the significant moderators from the full model, after dropping moderators due to multicollinearity until all moderators had a $GVIF^{1/(2*df)} < 2$. Then, we estimated an additional extension of the full model by adding all significant interactions that emerged from the previous interaction models. We next reduced that model until all moderators were significant at a 5% level. The resulting model achieved a higher AICc than our main reduced model. Comparing all full models with interactions, the model with the lowest AICc (Burnham and Anderson 2004) did not feature a significant interaction, indicating that the possible interactions are small and do not affect our results. All of these models are available in Web Appendix F.

Discussion

Theoretical contributions

Though three meta-analyses discussing the hypothetical bias exist (Carson et al. 1996; List and Gallet, 2001; Murphy et al. 2005), this is the first comprehensive study giving marketing managers and scholars advices on how to accurately measure consumers' WTP. In contrast to the existing meta-analyses, we focus on private goods, instead of on public goods, increasing the applicability of our findings within a marketing

Table 5 Hypotheses testing results

Hypothesis		Full model	Reduced model	Robustness checks
H1a	Type of measurement HWTP: indirect methods have smaller bias than direct methods			
H1b	Type of measurement HWTP: direct methods have smaller bias than indirect methods	✓	✓	✓
H2	Bias increases with product value	✓		✓
H3	Bias is least for convenience goods, greater for shopping goods, greatest for specialty goods	✓	✓	✓
H4	Bias is greater for innovations			
H5	Bias is greater for within-subject designs than for between-subject designs	✓		✓
H6	Opportunity to test a product reduces the bias			
H7a	Participation fee decreases the bias			
H7b	Initial balance decreases the bias			

context.[2] With a meta-analysis of 115 ESs gathered from 77 studies reported in 47 papers, we conclude that HWTP methods tend to overestimate RWTP considerably, by about 21% on average. This hypothetical bias depends on several factors, for which we formulated hypotheses (Table 5) and which we discuss subsequently.

With respect to the method for measuring HWTP, whether direct or indirect, across all the different models, we find strong support for H1b, which states that indirect methods overestimate HWTP more severely than direct methods. This important finding contradicts the prevailing opinion among academic researchers (Breidert et al. 2006) and has not previously been revealed in meta-analyses (Carson et al. 1996; List and Gallet 2001; Murphy et al. 2005). We in turn propose several potential mechanisms that could produce this surprising finding. First, we consider the concept of coherent arbitrariness, as first introduced by Ariely et al. (2003). People facing many consecutive choices tend to base each decision on their previous ones, such that they show stable preferences. However, study participants might make their first decision more or less randomly. Indirect measures require many, consecutive choices, so coherent arbitrariness could arise when using these methods to measure WTP. In that sense, the results of indirect measures indicate stable preferences, but they do not accurately reflect the participants' actual valuation. Second, participants providing indirect measure responses might focus less on the absolute values of an attribute and more on relative values (Drolet et al. 2000). The absolute values of the price attribute are key determinants of WTP, so the hypothetical bias might increase if the design of the choice alternatives does not include correct price levels. A widespread argument for the greater accuracy of indirect methods compared with direct methods asserts they mimic a natural shopping experience (Breidert et al. 2006); our analysis challenges this claim.

[2] Please refer to Web Appendix A for a more detailed discussion of the existing meta-analyses.

In our results related to H2, the *p* value of the *value* moderator is slightly greater than 5% in the full model, such that the hypothetical bias appears greater for more valuable products in percentage terms, though the effect is relatively small. *Value* does not remain in the reduced model, but the significant effect is very consistent across the robustness checks that feature the full model (Table 5). Therefore, our results support H2: The hypothetical bias increases if the value of the products to be evaluated increases. This finding is new, in that neither existing meta-analyses (Carson et al. 1996; List and Gallet 2001; Murphy et al. 2005) nor any primary studies have examined this moderating effect.

We also find support for H3 across all analyzed models. For participants it is harder to evaluate a specialty product's utility than a convenience product's utility; specialty goods often feature a higher degree of complexity or are less familiar to consumers than convenience goods. The greater ability to assess the product's utility reduces the hypothetical bias (Hofstetter et al. 2013), such that our finding of higher overestimation for specialty goods is in line with prior research. Yet we do not find any difference between shopping and convenience goods, prompting us to posit that the hypothetical bias might not be affected by moderate search effort; rather, only products demanding strong search effort increase the hypothetical bias. Existing meta-analyses (Carson et al. 1996; List and Gallet 2001; Murphy et al. 2005) include public goods and do not distinguish among different types of private goods. By showing that the type of a private good influences the hypothetical bias, we add to an understanding of the hypothetical bias in a marketing context that features private goods.

With respect to *innovation*, we find no support for H4, because the differences between innovations and existing products are small and not significant. This finding contrasts with Hofstetter et al.'s (2013) results. Accordingly, we avoid rejecting the claim that methods for measuring HWTP work as well (or as poorly) for innovations as they do for existing products.

A within-subject research design increases the hypothetical bias, compared with a between-subject design, as we predicted in H5 and in accordance with prior research (Ariely et al. 2006, Fox and Tversky 1995, Frederick and Fischhoff 1998). Yet this finding still seems surprising to some extent. When asking a participant for WTP twice (once hypothetically, once in a real context), the first answer seemingly should serve as an anchor for the second, leading to an assimilation expected to reduce the hypothetical bias. Instead, two similar questions under different conditions appear to evoke a contrast instead of an assimilation effect, and they produce a greater hypothetical bias. Consequently, when designing marketing experiments to investigate the hypothetical bias, researchers should use a between-subject design to prevent the answers from influencing each other. When researching the influence of consumer characteristics on the hypothetical bias though, it would be more appropriate to choose a within-subject design (Hofstetter et al. 2013), though researchers must recognize that the hypothetical bias might be overestimated more severely in this case. Murphy et al. (2005) also distinguish different subject designs in their meta-analysis and find a significant effect, though they use RWTP instead of the difference between HWTP and RWTP as their dependent variable. In this sense, our finding of a moderating role of the study design on the hypothetical bias is new to the literature.

Our results do not support H6; we do not find differences in the hypothetical bias when participants have an opportunity the test a product before stating their WTP or not. Testing a product in advance reduces uncertainty about product performance, and our finding is in contrast with Hofstetter et al.'s (2013) evidence that higher uncertainty increases the hypothetical bias. Note however, that the result by Hofstetter et al.'s (2013) refers to an effect of a consumer characteristic, and might be specific to the examined product, namely digital cameras. Our results are more general across a wide range of product categories and experimental designs. Furthermore, this result on H6 is in line with our findings for H4; both hypotheses rest on the participants' uncertainty about product performance, and we do not find support for either of them.

Finally, neither a *participation fee* nor *initial balance* reduce the hypothetical bias significantly, so we find no support for H7a or H7b. Formally, we can only "not reject" a null hypothesis of no moderator effect, but these findings suggest that we can dispel fears about influencing WTP results too much by offering participation fees or an initial balance.

In addition to these theoretical insights on WTP measures, we contribute to marketing literature by showing how to model stochastically dependent ESs explicitly when the covariances and variances of the observed ESs are known or can be computed. Moreover, we use (the log of) the response ratios as the ES in our meta-analysis, which has not been done previously in marketing. We provide a detailed rationale for using response ratios and thus offer marketing scholars another ES option to use in their meta-analyses.

Managerial implications

This meta-analysis identifies a substantial hypothetical bias of 21% on average in measures of WTP. Although hypothetically derived WTP estimates are often the best estimates available, managers should realize that they generally overestimate consumers' RWTP and take that bias into account when using HWTP results to develop a pricing strategy or when setting an innovation's launch price. In addition, we detail conditions in which the bias is larger or smaller, and we provide a brief overview of how extensive the expected biases might become. In particular, managers should anticipate a greater hypothetical bias when measuring WTP for products with higher values or for specialty goods. For example, when measuring HWTP for specialty goods, direct methods overestimate it by 28% and indirect methods do so by 40%. These predicted degrees of RWTP overestimation should be used to adjust decisions based on WTP studies in practice.

The study at hand also shows that direct methods result in more accurate estimates of WTP than indirect methods do. Therefore, practitioners can resist, or at least consider with some skepticism, the prevalent academic advice to use indirect methods to measure WTP. In addition to being less accurate, indirect methods require more effort and costs (Leigh et al. 1984). However, this recommendation only applies if the measurement of HWTP is necessary. If RWTP can be measured with an auction format, that option is preferable, since RWTP reflects actual WTP, whereas HWTP tends to overestimate it. This result also implies an exclusive focus on measuring WTP for a specific product, such that it disregards some advantages of the disaggregate information provided by indirect methods (e.g., demand due to cannibalization, brand switching, or market expansion; Jedidi and Jagpal 2009). In summary, the key takeaway for managers who might use direct measures of HWTP is that the "quick and dirty solution" is only quick, not dirty—or at least, not more dirty than indirect methods.

Limitations and research directions

This meta-analysis suggests several directions for further research, some of which are based on the limitations of our meta-analysis. First, several recent adaptations of indirect methods seek to improve their accuracy (Gensler et al. 2012, Schlereth and Skiera 2017). These improvements might reduce the variance in measurement accuracy between direct and indirect measurements. These recently developed methods have not been tested by empirical comparison studies, so we could not include them in our meta-analysis. An extensive comparison of those adaptions, in terms of their effects on the hypothetical bias, would provide researchers

and managers more comprehensive insights for choosing the right method when measuring WTP.

Second, the prevailing opinion of indirect methods yielding a lower hypothetical bias than direct methods bases upon assumptions concerning individuals' decision making; though our results are in contrast with this opinion. The underlying mental processes when asked for the WTP through direct or indirect methods are not well understood yet. Investigating those processes would foster the understanding of differences in the hypothetical bias between direct and indirect methods and between other experimental conditions. This would enable the development of new adaptions minimizing the hypothetical bias.

Third, the hypothetical bias depends on a variety of factors, including individual-level considerations (Hofstetter et al. 2013; Sichtmann et al. 2011), that extend beyond the product or study level moderators as examined in our meta-regressions. Very few studies have investigated these factors, so we could not incorporate them in our meta-analysis, though consumer characteristics likely explain some differences. Therefore, we call for more research on whether and how individual characteristics influence the hypothetical bias. For example, a possible explanation for the limited accuracy of indirect measures could reflect coherent arbitrariness (Ariely et al. 2003). Continued research might examine whether and how coherent arbitrariness affects different consumers, especially in the context of CBCs. In addition, our findings on some product-level factors are new, namely that the hypothetical bias is greater for higher valued products and for specialty goods. These results could be cross-validated in future experimental studies.

Fourth, knowing and measuring WTP is crucial for firms operating in business-to-business (B2B) contexts (Anderson et al. 1992), yet all ESs in our study are from a business-to-consumer context. Because B2B products and services tend to be more complex, customers might prefer to identify product characteristics and to include them separately when determining their WTP in response to an indirect method. However, anecdotal evidence indicates that direct measurement works better for industrial goods than for consumer goods (Dolan and Simon 1996). Researching the differential accuracy of the various methods in a B2B context would be especially interesting; our study already indicates differences between convenience and (more complex) specialty goods. Therefore, we join Lilien (2016) in calling for more research in B2B marketing, including the measurement of WTP.

Fifth, the majority of studies included herein used open questioning as the direct method for measuring WTP. In practice, different direct methods are available (Steiner and Hendus 2012), yet they rarely have been investigated in academic research. Pricing research could increase in managerial relevance (Borah et al. 2018), and help managers make better pricing decisions, if it included assessments of different direct methods for measuring WTP.

Acknowledgments The authors appreciate helpful comments from Felix Eggers, Manfred Krafft, and Hans Risselada.

References

Abraham, A. T., & Hamilton, R. W. (2018). When does partitioned pricing lead to more favorable consumer preferences? Meta-analytic evidence. *Journal of Marketing Research, 55*(5), 686–703.

Anderson, J. C., Jain, D. C., & Chintagunta, P. K. (1992). Customer value assessment in business markets: A state-of-practice study. *Journal of Business-to-Business Marketing, 1*(1), 3–29.

Ariely, D., Loewenstein, G., & Prelec, D. (2003). "Coherent arbitrariness": Stable demand curves without stable preferences. *The Quarterly Journal of Economics, 118*(1), 73–106.

Ariely, D., Ockenfels, A., & Roth, A. E. (2005). An experimental analysis of ending rules in internet auctions. *RAND Journal of Economics, 36*(4), 890–907.

Ariely, D., Loewenstein, G., & Prelec, D. (2006). Tom sawyer and the construction of value. *Journal of Economic Behavior & Organization, 60*(1), 1–10.

Arts, J. W., Frambach, R. T., & Bijmolt, T. H. A. (2011). Generalizations on consumer innovation adoption: A meta-analysis on drivers of intention and behavior. *International Journal of Research in Marketing, 28*(2), 134–144.

Babić Rosario, A., Sotgiu, F., de Valck, K., & Bijmolt, T. H. A. (2016). The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors. *Journal of Marketing Research, 53*(3), 297–318.

Barrot, C., Albers, S., Skiera, B., & Schäfers, B. (2010). Why second-price sealed-bid auction leads to more realistic price-demand functions. *International Journal of Electronic Commerce, 14*(4), 7–38.

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

Bijmolt, T. H. A., & Pieters, R. G. M. (2001). Meta-analysis in marketing when studies contain multiple measurements. *Marketing Letters, 12*(2), 157–169.

Bijmolt, T. H. A., van Heerde, H. J., & Pieters, R. G. M. (2005). New empirical generalizations on the determinants of price elasticity. *Journal of Marketing Research, 42*(2), 141–156.

Bolton, G. E., & Ockenfels, A. (2014). Does laboratory trading mirror behavior in real world markets? Fair bargaining and competitive bidding on eBay. *Journal of Economic Behavior & Organization, 97*, 143–154.

Borah, A., Wang, X., & Ryoo, J. H. (2018). Understanding influence of marketing thought on practice: An analysis of business journals using textual and latent Dirichlet allocation (LDA) analysis. *Customer Needs and Solutions, 5*(3–4), 146–161.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, United Kingdom: John Wiley & Sons.

Breidert, C., Hahsler, M., & Reutterer, T. (2006). A review of methods for measuring willingness-to-pay. *Innovative Marketing, 2*(4), 8–32.

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.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*(2), 261–304.

Bushong, B., King, L. M., Camerer, C. F., & Rangel, A. (2010). Pavlovian processes in consumer choice: The physical presence of a good increases willingness-to-pay. *American Economic Review, 100*(4), 1556–1571.

Carson, R. T., Flores, N. E., Martin, K. M., & Wright, J. L. (1996). Contingent valuation and revealed preference methodologies: Comparing the estimates for quasi-public goods. *Land Economics, 72*(1), 80–99.

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within subject designs. *Journal of Economic Behavior & Organization, 81*(1), 1–8.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Copeland, M. T. (1923). Relation of consumers' buying habits to marketing methods. *Harvard Business Review, 1*(3), 282–289.

Dimoka, A., Hong, Y., & Pavlou, P. A. (2012). On product uncertainty in online markets: Theory and evidence. *MIS Quarterly, 36*(2), 395–426.

Ding, M. (2007). An incentive-aligned mechanism for conjoint analysis. *Journal of Marketing Research, 44*(2), 214–223.

Ding, M., Grewal, R., & Liechty, J. (2005). Incentive-aligned conjoint analysis. *Journal of Marketing Research, 42*(1), 67–82.

Dolan, R. J., & Simon, H. (1996). *Power pricing: how managing price transforms the bottom line.* New York: The Free Press.

Drolet, A., Simonson, I., & Tversky, A. (2000). Indifference curves that travel with the choice set. *Marketing Letters, 11*(3), 199.209.

Edeling, A., & Fischer, M. (2016). Marketing's impact on firm value: Generalizations from a meta-analysis. *Journal of Marketing Research, 53*(4), 515–534.

Edeling, A., & Himme, A. (2018). When does market share matter? New empirical generalizations from a meta-analysis of the market share–performance relationship. *Journal of Marketing, 82*(3), 1–24.

Eggers, F., & Sattler, H. (2009). Hybrid individualized two-level choice-based conjoint (HIT-CBC): A new method for measuring preference structures with many attribute levels. *International Journal of Research in Marketing, 26*(2), 108–118.

Fox, J., & Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association, 87*(417), 178–183.

Fox, C. R., & Tversky, A. (1995). Ambiguity aversion and comparative ignorance. *The Quarterly Journal of Economics, 110*(3), 585–603.

Frederick, S., & Fischhoff, B. (1998). Scope (in)sensitivity in elicited valuations. *Risk Decision and Policy, 3*(2), 109–123.

Gensler, S., Hinz, O., Skiera, B., & Theysohn, S. (2012). Willingness-to-pay estimation with choice-based conjoint analysis: Addressing extreme response behavior with individually adapted designs. *European Journal of Operational Research, 219*(2), 368–378.

Gensler, S., Neslin, S. A., & Verhoef, P. C. (2017). The showrooming phenomenon: It's more than just about price. *Journal of Interactive Marketing, 38*, 29–43.

Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 357–376). New York: Russel Sage Foundation.

Grewal, D., Puccinelli, N., & Monroe, K. B. (2017). Meta-analysis: Integrating accumulated knowledge. *Journal of the Academy of Marketing Science, 47*(5), 840.

Hair J.F. Jr, Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis (8th ed.).* Hampshire, United Kingdom: Cengage Learning EMEA.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorial in Quantitative Methods for Psychology, 8*(1), 23–34.

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* (Vol. 1, pp. 752–767). Amsterdam, Netherlands: Elsevier.

Hedges, L. V., Gurevitch, J., & Curtis, P. S. (1999). The meta-analysis of response ratios in experimental ecology. *Ecology, 80*(4), 1150–1156.

Hensher, D. A. (2010). Hypothetical bias, choice experiments and willingness to pay. *Transportation Research Part B: Methodological, 44*(6), 735–752.

Hoeffler, S. (2003). Measuring preferences for really new products. *Journal of Marketing Research, 40*(4), 406–420.

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.

Ingenbleek, P. T. M., Frambach, R. T., & Verhallen, T. M. M. (2013). Best practices for new product pricing: Impact on market performance and price level under different conditions. *Journal of Product Innovation Management, 30*(3), 560–573.

Jedidi, K., & Jagpal, S. (2009). Willingness to pay: Measurement and managerial implications. In V. R. Rao (Eds.), *Handbook of pricing research in marketing* (pp. 37–60). Cheltenham, United Kingdom: Edward Elgar Publishing.

Jedidi, K., & Zhang, Z. J. (2002). Augmenting conjoint analysis to estimate consumer reservation price. *Management Science, 48*(10), 1350–1368.

Kagel, J. H., Harstad, R. M., & Levin, D. (1987). Information impact and allocation rules in auctions with affiliated private values: A laboratory study. *Econometrica, 55*(6), 1275–1304.

Kalaian, H. A., & Raudenbush, S. W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods, 1*(3), 227–235.

Kimenju, S. C., Morawetz, U. B., & De Groote, H. (2005). Comparing contingent valuation method, choice experiments and experimental auctions in soliciting consumer preference for maize in Western Kenya: Preliminary results (Presentation at the African Econometric Society 10th annual conference on econometric modeling in Africa, Nairobi, Kenya).

Kohli, R., & Mahajan, V. (1991). A reservation-price model for optimal pricing of multiattribute products in conjoint analysis. *Journal of Marketing Research, 28*(3), 347–354.

Koricheva, J., & Gurevitch, J. (2014). Uses and misuses of meta-analysis in plant ecology. *Journal of Ecology, 102*(4), 828–844.

Lajeunesse, M. J. (2011). On the meta-analysis of response ratios for studies with correlated and multi-group designs. *Ecology, 92*(11), 2049–2055.

Leeflang, P.S.H., Wieringa, J.E., Bijmolt, T.H.A., & Pauwels, K.H. (2015). Modeling markets; analyzing marketing phenomena and improving marketing decision making. New York, NY: Springer.

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.

Lilien, G. (2016). L. (2016). The b2b knowledge gap. *International Journal of Research in Marketing, 33*, 543–556.

List, J. A., & Gallet, C. A. (2001). What experimental protocol influence disparities between actual and hypothetical stated values? Evidence from a meta-analysis. *Environmental and Resource Economics, 20*(3), 241–254.

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), 172–184.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7*(1), 105–125.

Murphy, J. J., Allen, P. G., Stevens, T. H., & Weatherhead, D. (2005). A meta-analysis of hypothetical bias in stated preference valuation. *Environmental and Resource Economics, 30*(3), 313–325.

Nagle, T. T., & Müller, G. (2018). *The strategy and tactics of pricing: A guide to growing more profitably* (6th ed.). New York, NY: Routledge.

Neill, H. R., Cummings, R. G., Ganderton, P. T., Harrison, G. W., & McGuckin, T. (1994). Hypothetical surveys and real economic commitments. *Land Economics, 70*(2), 145–154.

Noussair, C., Robin, S., & Ruffieux, B. (2004). Revealing consumers' willingness-to-pay: A comparison of the BDM mechanism and the Vickrey auction. *Journal of Economic Psychology, 25*(6), 725–741.

Ockenfels, A., & Roth, A. E. (2006). Late and multiple bidding in second price internet auctions: Theory and evidence concerning different rules for ending an auction. *Games and Economic Behavior, 55*(2), 297–320.

Pebsworth, P. A., MacIntosh, A. J. J., Morgan, H. R., & Huffman, M. A. (2012). Factors influencing the ranging behavior of chacma baboons (papio hamadryas ursinus) living in a human-modified habitat. *International Journal of Primatology, 33*(4), 872–887.

Rutström, E. E. (1998). Home-grown values and incentive compatible auction design. *International Journal of Game Theory, 27*(3), 427–441.

Scheibehenne, B., Greifeneder, R., & Todd, P. M. (2010). Can there ever be too many options? A meta-analytic overview of choice overload. *Journal of Consumer Research, 37*(3), 409–425.

Schlag, N. (2008). Validierung der Conjoint-Analyse zur Prognose von Preisreaktionen mithilfe realer Zahlungsbereitschaften. In *Lohmar*. Germany: Josef Eul Verlag.

Schlereth, C., & Skiera, B. (2017). Two new features in discrete choice experiments to improve willingness-to-pay estimation that result in SDR and SADR: Separated (adaptive) dual response. *Management Science, 63*(3), 829–842.

Shogren, J. F., Margolis, M., Koo, C., & List, J. A. (2001). A random nth-price auction. *Journal of Economic Behavior & Organization, 46*(4), 409–421.

Sichtmann, C., Wilken, R., & Diamantopoulos, A. (2011). Estimating willingness-to-pay with choice-based conjoint analysis: Can consumer characteristics explain variations in accuracy? *British Journal of Management, 22*(4), 628–645.

Simon, H. (2018). Irrationals Verhalten. Interview. *Harvard Business Manager, 40*(8), 52–54.

Steiner, M., & Hendus, J. (2012). How consumers' willingness to pay is measured in practice: An empirical analysis of common approaches' relevance. Retrieved from SSRN: https://ssrn.com/abstract=2025618. Accessed 20 Aug 2018

Steiner, M., Eggert, A., Ulaga, W., & Backhaus, K. (2016). Do customized service packages impede value capture in industrial markets? *Journal of the Academy of Marketing Science, 44*(2), 151–165.

Thompson, S. G., & Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: A comparison of methods. *Statistics in Medicine, 18*(20), 2693–2708.

Tully, S. M., & Winer, R. S. (2014). The role of the beneficiary in willingness to pay for socially responsible products: a meta-analysis. *Journal of Retailing, 90*(2), 255–274.

van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods, 45*(2), 576–594.

van Houwelingen, H. C., Arends, L. R., & Stijnen, T. (2002). Advanced methods in meta-analysis: Multivariate approach and meta-regression. *Statistics in Medicine, 21*(4), 589–624.

Vega, L. A., Koike, F., & Suzuki, M. (2010). Conservation study of myrsine seguinii in Japan: Current distribution explained by past land use and prediction of distribution by land use-planning simulation. *Ecological Research, 25*(6), 1091–1099.

Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance, 16*(1), 8–37.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3).

Völckner, F. (2006). Methoden zur Messung individueller Zahlungsbereitschaften: Ein Überblick zum State of the Art. *Journal für Betriebswirtschaft, 56*(1), 33–60.

Wang, T., Venkatesh, R., & Chatterjee, R. (2007). Reservation price as a range: An incentive-compatible measurement approach. *Journal of Marketing Research, 44*(2), 200–213.

Wertenbroch, K., & Skiera, B. (2002). Measuring consumers' willingness to pay at the point of purchase. *Journal of Marketing Research, 39*(2), 228–241.

Wlömert, N., & Eggers, F. (2016). Predicting new service adoption with conjoint analysis: External validity of BDM-based incentive-aligned and dual-response choice designs. *Marketing Letters, 27*(1), 195–210.