WORKING PAPER · NO. 2022-35

Improving Willingness-to-Pay Elicitation by Including a Benchmark Good

Rebecca Dizon-Ross and Seema Jayachandran

MARCH 2022

Becker Friedman Institute
FOR ECONOMICS AT UCHICAGO

5757 S. University Ave.
Chicago, IL 60637
Main: 773.702.5599
bfi.uchicago.edu

IMPROVING WILLINGNESS-TO-PAY ELICITATION BY INCLUDING A BENCHMARK GOOD

Rebecca Dizon-Ross
Seema Jayachandran

March 2022

We thank Rebekah Chang and Suanna Oh for excellent research assistance and Kate Orkin for helpful comments. Jayachandran gratefully acknowledges support from the National Science Foundation (SES-1156941). We also thank the Equality Development and Globalization Studies initiative at Northwestern University for funding. This project received IRB approval from Northwestern University, Stanford University, and the Uganda National Council for Science and Technology.

Improving Willingness-to-Pay Elicitation by Including a Benchmark Good
Rebecca Dizon-Ross and Seema Jayachandran
March 2022
JEL No. C83,O1

ABSTRACT

We propose and validate a simple way to augment the standard Becker-DeGroot-Marschak method that researchers use to elicit willingness to pay (WTP) for a good. The augmentation is to measure WTP for another good ("benchmark good"), one unrelated to both the good the researcher is interested in and the independent variables of interest, and to use WTP for the benchmark good as a control variable in analyses. We illustrate the method and how it can eliminate noise in measured WTP using data collected in Uganda.

Rebecca Dizon-Ross
Booth School of Business
University of Chicago
5807 South Woodlawn Avenue
Chicago, IL 60637
and NBER
rdr@chicagobooth.edu

Seema Jayachandran
Department of Economics
Northwestern University
2211 Campus Dr
Evanston, IL 60208
and NBER
seema@northwestern.edu

1 Introduction

In this article, we propose and validate a simple way to augment the standard procedure that researchers use to elicit willingness to pay (WTP) for a good or service. The objective is to improve the accuracy and precision of measured WTP by eliminating some of the variation in raw responses that is unrelated to true WTP.

The context for the validation exercise is a study in Uganda about parents' preferences for goods for their children. To understand these preferences, we elicited WTP for various goods for children using the Becker-DeGroot-Marschak (BDM) method (Becker, DeGroot and Marschak, 1964). In the most common version of BDM, the respondent gives their take-it-or-leave-it purchase decision at several prices, knowing that one of the prices will be chosen at random and a transaction will or will not take place at that price, depending on their response. The respondent's weakly dominant strategy is to answer truthfully.

The augmentation we propose is to measure WTP for a "benchmark good" that is unrelated to both the good the researcher is interested in and the independent variables of interest. WTP for the benchmark good is then used as a control variable in analyses (or, more generally, to adjust WTP for the focal goods). For example, in our context, the focal goods were human capital goods for children. We were interested in how child gender affects WTP for human capital goods, as well as how mothers' and fathers' WTP differ. We used household goods that were not systematically preferred more by men or women, such as containers for water, as the benchmark goods. WTP for water containers is also unrelated to how much one values human capital for a girl versus a boy, in all likelihood. Another consideration when choosing benchmark goods is the income elasticity of demand for them, which we discuss at the end of the article.

Below, we describe the study setting and data and show evidence of systematic noise in elicited WTP (e.g., day-of-the-week effects). We then demonstrate how controlling for the benchmark good WTP can reduce this noise. We conclude by discussing when this approach would help researchers answer their research question more effectively and when it might instead constitute "over-controlling."

2 Study sample and data collection

We collected data in Iganga district in Uganda among a sample of two-parent households with a child in primary school. We collaborated with local schools to obtain contact information for eligible households in early 2013. We visited households to verify eligibility and enrolled 1,084 of them in the study. We randomly selected either the mother or father for the survey. The first survey took place in May 2013.

We made a second household visit in September to October 2013 to interview the other parent. We made these second visits to 729 households, focusing on those that also had a child age 3 to 8 years old. The reason for the original criterion of having a child in primary school and the follow-up criterion of having a younger child was that we elicited WTP for goods that were relevant for children meeting these criteria.

The data were collected for a study on parental preferences related to their children in which we pool the WTP data from both visits (Dizon-Ross and Jayachandran, 2022). However, in this article, we focus on the data from the second survey because one of the key variables we use (the respondent's propensity to give socially desirable answers) was not collected in the first survey. Because we randomized which parent was surveyed in the first wave and interviewed the other parent in the second wave, the sample is balanced by gender; 49% of respondents in the second wave are female.

The survey included several questions about socioeconomic background and views regarding children. We also included the Marlowe-Crowne module, which was developed by social psychologists to measure a person's propensity to give socially desirable answers (Crowne and Marlowe, 1960). The module asks respondents if they have several exemplary traits such as never being jealous of another person's good fortune. The module uses near-saintly traits that most people do not truly possess. The premise is that people who report having more of these traits likely have a stronger concern for social approval or a stronger desire to look good to themselves. We included a 13-question version of the module following Dhar, Jain and Jayachandran (2022). We calculate the proportion of exemplary traits that the respondent reports possessing and use it as a measure of the respondent's propensity to give socially desirable answers.

Most of the survey was focused on eliciting WTP for several goods using BDM. We used a multiple price list, asking the respondent if they were willing to purchase the good at a

series of prices in descending order from the market price to a price near 0. The decrement varied by good and was chosen so that respondents were asked about roughly 12 price levels per good. We told participant the market price before starting the elicitation. While this prevented us from eliciting WTP above the market price, in piloting we found that this helped respondents understand BDM. Accordingly, we chose goods that few participants would purchase at an unsubsidized price.

The respondent was told that after the price questions, one of the prices would be randomly chosen, with equal probability, and they would purchase the good from us at that price if and only if their answer had been that they wanted to. The selection of the randomized price (done via tablet) and exchange of money and goods, if applicable, was conducted just after the BDM questions were asked for a good.

BDM seemed to work well in our study. While the exercise was initially unfamiliar to participants, we explained the procedure to them in detail and resolved any confusion. After the actual BDM was conducted, we asked debrief questions such as regret about one's choices to confirm understanding. Based on these responses, comprehension was high.

We chose the goods carefully because, in preliminary fieldwork, we found that certain criteria were key to BDM working well. First, it helped if people were familiar with the good and its market price; otherwise, variation in perceived quality or beliefs about whether the good could be purchased elsewhere for less than the market price we indicated would add noise. Second, we used goods that most respondents valued at less than the market price, but placed some value on, so that there was variation in WTP. There were also practical considerations such as the good being small and light enough for surveyors to carry around on foot.

All respondents in the second wave had both a primary-school-age (hereafter, older) child and a 3-to-8 year old (hereafter, younger) child. We used BDM to elicit WTP for a math workbook and rubber ball for the older child and rubber shoes and a bag of candy for the younger child. To increase sample size, we also used similar price lists to elicit WTP in a non-incentivized way for three goods: practice exams for the older child and deworming medicine and a toy for the younger child. For the non-incentivized goods, the surveyor showed the actual good to the respondent so that it was concrete, but respondents knew in advance that for these goods, no transaction would take place. The WTP elicitation performed well for these goods too, perhaps because they were embedded in a survey that

also conducted incentivized elicitation in a similar format.

In the analysis, we pool the incentivized and non-incentivized goods (seven goods in total). We show that the results are very similar if we include only the incentivized goods. The goods ranged in price from 2,000 to 4,500 Ugandan shillings (UGX), where 1 US dollar $\approx$ 2,600 Ugandan shillings at the time.

We also elicited WTP (in an incentivized way) for two goods for adults that serve as benchmark goods. The purpose was to use WTP for them as a control variable to address possible noise from misunderstanding of the BDM procedure, short-term fluctuations in cash-on-hand, gift-exchange behavior, etc. We made one of these benchmark goods the first BDM exercise, in part to get gift exchange out of the respondent's system, so to speak. The two benchmark goods were a poster of the Uganda parliament and a pair of plastic jerry cans (large containers for liquid).

The main outcome data are WTP measures, and to make them comparable across goods, we divide WTP for each good by its sample mean. To combine WTP for the two benchmark goods into one control variable, we divide each variable by its sample mean and then average them. One concern with including the benchmark goods is that it might exacerbate the role of cash on hand. All participants received 9,000 UGX upfront for participation, more than the maximum price of the benchmark goods.

3 Analyses to illustrate and fix measurement error in WTP

We now document spurious variation in elicited WTP and test whether controlling for the benchmark good WTP helps address the problem. We focus on two factors: whether the survey takes place on the weekend and the respondent's propensity for social desirability bias.

3.1 Weekend effects

The work week for the survey team was Monday to Saturday, with a small amount of additional surveying conducted on Sunday; 15% of the sample surveyed on Saturday or Sunday. Table 1, column 1 reports how WTP for children's goods differs when the survey was on the weekend. The regression pools the seven children's goods so has 5,103 observations (729 respondents $\times$ 7 goods). Standard errors are clustered at the respondent level.

Table 1: WTP differs on weekends

	Child goods WTP (1)	Benchmark goods WTP (2)	Child goods WTP (3)
Weekend	-0.139	-0.170	-0.024
	[0.055]	[0.055]	[0.034]
Benchmark goods WTP			0.680
			[0.024]
Observations	5103	729	5103

Note: Each observation is a person-good. Standard errors are clustered by person.

We find that WTP for children's goods is significantly lower among those surveyed on weekends (p-value = 0.01). WTP is normalized to have an average value of 1 for each good, so the coefficient of -0.14 corresponds to roughly 14% lower WTP. The results are very similar if we exclude surveys conducted on Sunday.

The weekend effect could arise for various reasons. The composition of who is surveyed might differ or the survey conditions might differ. However, when we conduct a balance check of characteristics of the respondent and the survey environment (e.g., gender, household asset index, polygamous household, whether the survey was conducted at home), we find that characteristics are balanced.

Another possibility is that people have less cash on hand because they spent it on Friday or their pay day is early in the week. This also does not seem to be the explanation, as there is no significant correlation between cash on hand and weekend surveys.

Perhaps people were more or less tired or in a better or worse mood on weekends. The fact that the obvious explanations such as sample composition and cash on hand are not the explanation makes directly controlling for the root cause challenging.

The benchmark goods are useful for addressing this problem because, as seen in Table 1, column 2, the same pattern of a lower WTP on the weekend holds for them (p-value $<$ 0.01). (For simplicity, we averaged the WTP for the two benchmark goods, but the results are similar if we use them separately.) Column 3, shows that when children's good WTP is the outcome, controlling for the benchmark good WTP knocks out most of the weekend effect.

Table 2: WTP varies with respondent's propensity for social desirability bias

	Child goods WTP (1)	Benchmark goods WTP (2)	Child goods WTP (3)
Social desirability score	-0.225	-0.350	0.014
	[0.132]	[0.133]	[0.085]
Benchmark goods WTP			0.683
			[0.024]
Observations	5103	729	5103

Note: Each observation is a person-good. Standard errors are clustered by person.

3.2 Social desirability bias

We next analyze how WTP varies with the respondent's propensity for social desirability bias. Recall that the measure is the proportion of self-reported exemplary traits, so the variable can take on values from 0 to 1. In our sample, the median is 0.69 and the 25th and 75th percentiles are 0.54 and 0.77, respectively.

The measure indeed seems to predict a respondent's desire to look good. For example, people with a higher social desirability bias (SDB) score are significantly more likely to say that their older child is very likely to continue in school next year.

Our prior was that respondents with a higher SDB score might report a higher WTP because they wanted to seem generous to the surveyor. Instead, we find quite consistently that a higher SDB score is associated with a lower WTP. Perhaps people want to portray themselves or maintain a self-identity as thrifty, or people who did not fully understand BDM wanted to come across as aggressive bargainers.[1]

Table 2, column 1, reports the negative effect of the SDB score on WTP for children's good (p-value = 0.09). This effect size is modest: Moving from the 25th to the 75th percentile of the SDB score distribution is associated with 5% lower WTP. Column 2 shows a similar, and in fact stronger, pattern for the benchmark goods (p-value < 0.01). Finally, column 3 shows that controlling for the benchmark good WTP when analyzing WTP for the children's goods shrinks the coefficient on the SDB score to close to zero.

[1]We find no evidence that the SDB score is strongly correlated with other traits that predict true demand. Its correlation with the respondent's gender, education level, cash on hand, and household asset index all have an absolute value smaller than 0.02.

Table 3: Excluding goods with non-incentivized WTP elicitation

	Child goods WTP (1)	Child goods WTP (2)
Weekend	-0.132	-0.018
	[0.052]	[0.033]
Social desirability score	-0.197	0.031
	[0.123]	[0.077]
Benchmark goods WTP		0.713
		[0.022]
Observations	2916	2916

Note: Each observation is a person-good. Standard errors are clustered by person.

3.3 Excluding non-incentivized goods

The analyses above pool the four goods that use standard BDM and the three goods for which we elicited WTP in a parallel but non-incentivized way. As a robustness check, Table 3 restricts the sample to the incentivized goods. The patterns seen earlier are also seen for the subsample of children's goods with incentivized price elicitation (column 1), and controlling for the benchmark good WTP eliminates most of this noise (column 2).

As a side note, even beyond this robustness check, we observe quite similar properties for the incentivized and non-incentivized goods. The fact that the stated preference WTP seems to provide meaningful data — at least when embedded in a standard incentivized BDM exercise — suggests that adding such hypothetical choices could be a fruitful way to increase sample size in applications of BDM. The advantages are time and logistical savings due to not having to execute the transactions, plus cost savings from not having to subsidize as many purchases.

4 Discussion

The analyses above document noise in WTP: the day of the week and the respondent's tendency to give socially desirable survey answers are systematically associated with WTP. These components of measured WTP do not reflect the persistent, true WTP of interest, so removing them is desirable. We show that that this can be done by using the WTP for a benchmark good as a control variable.

That said, whether variation is "true" or "just noise" is a conceptually subtle question.

If a person's desire to seem frugal influences their purchasing behavior beyond the study, that variation in WTP might be desirable to keep.

Moreover, measured WTP for the benchmark good reflects both noise and true preferences for the good. Thus, it is important to choose benchmark goods that are unrelated to the focal goods. Specifically, preferences for the two types of goods should be orthogonal, like preferences for jerry cans and children's goods seemed to be.

However, budget constraints matter too. Demand for most goods exhibits income effects, and controlling for the benchmark good WTP would be over-controlling if it eliminates these income effects.

One way to address this is to choose inexpensive benchmark goods, or more generally goods with weak income effects on demand. In our application, WTP for the benchmark goods was more strongly associated with temporary income, as proxied by cash on hand, than with permanent income, as proxied by a household asset index. In many cases, researchers are interested in the persistent component of demand, so purging temporary income effects is desirable.

Finally, sometimes conditioning out persistent income effects is a feature, not a bug. Dizon-Ross and Jayachandran (2022) examine how mothers' and fathers' spending preferences differ. For that research question, within-couple variation in income is a confound. Thus, an additional benefit of using benchmark goods is that they help control for income effects, resulting in a more precise and accurate measure of the object of interest.

References

Becker, Gordon M, Morris H DeGroot, and Jacob Marschak. 1964. "Measuring Utility by a Single-Response Sequential Method." *Behavioral Science*, 9(3): 226–232.

Crowne, Douglas P., and David Marlowe. 1960. "A New Scale of Social Desirability Independent of Psychopathology." *Journal of Consulting Psychology*, 24(4): 349–354.

Dhar, Diva, Tarun Jain, and Seema Jayachandran. 2022. "Reshaping Adolescents' Gender Attitudes: Evidence from a School-Based Experiment in India." *American Economic Review*, 112(3).

Dizon-Ross, Rebecca, and Seema Jayachandran. 2022. "Dads and Daughters: Disentangling Altruism and Investment Motives for Spending on Children." Working paper, University of Chicago.