PRELIMINARY AND INCOMPLETE: PLEASE DO NOT CITE OR CIRCULATE WITHOUT PERMISSION

Using Massive Online Choice Experiments to Measure Changes in Well-being

Erik Brynjolfsson
Massachusetts Institute of Technology and NBER

Felix Eggers
University of Groningen

Avinash Gannamaneni
Massachusetts Institute of Technology

February 23, 2017

Abstract

GDP and metrics derived from it, like productivity, have been central to our understanding of economic progress. In principle, changes in consumer surplus (compensating expenditure) provide a superior measure of changes in consumer well-being, especially for digital goods. In practice, consumer surplus has been difficult to measure. We explore the potential for massively scalable online Single Binary Discrete Choice experiments. These experiments seek to measure consumers' willingness to accept compensation for losing access to various digital goods and thereby estimate the changes in consumer surplus from these goods. Our results indicate that digital goods have created enormous gains in well-being which are largely missed by conventional measure of GDP and productivity, and suggest that our approach can be scaled up to a broader set of goods and services. Two limitations of our methods are that they are much less precise than changes in GDP and they suffer from hypothetical bias. We show how much of an improvement in precision can be achieved with larger sample sizes and various demographic controls and we document the direction and magnitude of bias present in our approach by conducting incentive compatible studies. By periodically querying a large, representative sample of goods and services, including those which are not priced in existing markets, changes in consumer surplus and other new measures of well-being derived from these online choice experiments have the potential for providing cost-effective supplements to existing national income and product accounts.

Keywords: Consumer Surplus, Digital Goods, Free Goods, GDP, Choice Experiments, Well-being

JEL Classification: C82

Acknowledgements: We thank Susanto Basu, Carol Corrado, Erwin Diewert, Robert Hall, John Hauser, Charles Hulten, Dale Jorgenson, Adam Saunders, Dan Sichel, Hal Varian and participants of the NBER Summer workshop for the Conference on Research on Income and Wealth (CRIW 2016), Micro@Sloan conference (2016) and Workshop on Information Systems and Economics (2016) for helpful comments and the MIT Initiative on the Digital Economy, through a grant from the Markle Foundation, for generous funding.

"If you don't know where you're going, you might not get there." -- Yogi Berra

1. Introduction

Digital technologies have transformed the nature of production and the types of goods and services consumed in modern economies. Yet our measurement framework for economic growth and well-being has not fundamentally changed since the 1930s. In principle, a more comprehensive approach is now feasible. By using massive online choice experiments to estimate changes in consumer surplus (compensating variation) we can supplement the traditional GDP-based metrics.

Gross Domestic Product (GDP) measures the monetary value of the purchases of all final goods by households, businesses and government. It is the most widely used measure of economic activity and heavily influences policymakers in setting economic objectives. GDP is heralded as one of the greatest inventions of the 20th century by noted economists Paul Samuelson and William Nordhaus (Landefeld 2000). However many economists agree that GDP is a significantly flawed measure of well-being and attempts are being made to design alternative measures (Stiglitz et al. 2009).

Both economists and journalists routinely use GDP as if it were a welfare measure. Media articles regularly mention that the "economy grew by x%"[1] by measuring the growth in GDP, and use this figure as a casual metric for the improvement in economic well-being. Similarly, economists widely use GDP per hour worked as a measure of productivity and derive

[1] E.g. U.S. Economy Grew 1.4% in Fourth Quarter (http://www.bloomberg.com/news/articles/2016-03-25/u-s-economy-grew-1-4-in-fourth-quarter-supported-by-consumers), China's Economy Grew by 6.7% in First Quarter of 2016 (http://blogs.wsj.com/chinarealtime/2016/04/15/chinas-economy-grew-by-6-7-in-first-quarter-of-2016/)

links between productivity and improvement in living standards (OECD 2008). However many economists have long acknowledged that GDP is not a welfare measure. In fact, Simon Kuznets, the founding father of the system of national accounts that include GDP, explicitly warned against using it this way, writing "The welfare of a nation can scarcely be inferred from a measurement of national income as defined [by the GDP.]" (Kuznets 1934).[2] Despite Kuznets' warning, growth in GDP is still the most widely used indicator of economic progress.

GDP as a welfare measure is especially problematic in the emerging digital economy since most of the digital goods have nearly zero marginal cost and often a zero equilibrium price. Most of their contributions are therefore invisible in the GDP calculations (Brynjolfsson and Saunders 2009; Brynjolfsson and McAfee 2014). For instance, although information goods have become increasingly more ubiquitous and important in our daily lives, the share of the information sector as a fraction of the total GDP (~ 4-5%) has not changed in the last 35 years.[3] Moreover, in many sectors (e.g. music, media, encyclopedias) people substitute paid (offline) goods with free online services (e.g., Spotify, YouTube, Wikipedia) so that the total revenue that shows up in GDP figures could fall even while consumers get access to better quality and more variety of digital goods (Brynjolfsson and Saunders 2009). In other words, not only the magnitude, but even the sign of the change in well-being may be incorrectly inferred if decision makers rely on existing measures of GDP and productivity as a proxy for well-being.

The benefits of technological advance are distinct from the expenditures on goods and services. Nordhaus (2005) estimated that between 1948 and 2001 corporations were able to

---

[2] He underscored his views when accepting his Nobel Prize in 1971, saying that the conventional measures of national product (including GDP) omit various costs (e.g. pollution) and benefits (e.g. more leisure time) associated with technological innovations and predicted major changes in the way we measure the economy (Kuznets 1973).
[3] https://www.bea.gov/industry/gdpbyind_data.htm

retain only 3.7% of the social returns from their technological advances while the remaining 96.3% of social returns went to consumers. Consumer surplus[4] thus reflects most of the returns to improvements in technology.

So far, the change in consumer surplus hasn't been widely used as a measure of well-being not because it is a poor measure of well-being, but because it is difficult to measure at scale. Estimating demand curves requires exogenous variations that shift the supply curve but not the demand curve and it is typically not practical to identify these variations for a large bundle of goods. However, with advancements in digital technologies it is now feasible to collect data about thousands of goods easily and quickly. These big data techniques have the potential to improve measurement of economic indicators. For example, Cavallo and Rigobon (2016) scrape the web to collect billions of prices for millions of products to construct prices indices and inflation measures for various countries as part of MIT's Billion Prices Project. Private companies like Amazon, Google and Facebook routinely conduct millions of online experiments to help understand consumer preferences and behavior. This scale of experimentation and inference would have been infeasible 20 years ago, but is now routine, at least in private industry.

In this research, we propose a way of measuring the changes in consumer surplus, first for the digital economy and then more broadly. We assess the feasibility of this approach with a small sample of goods. Through thousands of discrete choice experiments conducted online, we measure consumers' willingness to accept for losing access to various digital goods for a representative sample of the US population and estimate the demand curves for these goods. We conclude that our approach is easily scalable and can be used to develop a system that tracks

[4] Consumer surplus can be defined as the difference between a consumer's maximum willingness to pay for a product and the actual price paid for it.

changes in consumer surplus of numerous goods and services in (near) real time via massive online choice experiments.

The paper proceeds as follows. In section 2, we illustrate the ways the GDP and consumer surplus change when prices change or new products are introduced, and the implications for welfare estimates. Section 3 describes the methodology we use to empirically assess consumer surplus. Section 4 provides the key findings and Section 5 concludes with a summary and brief discussion.

2. Background

2.1 GDP, consumer surplus and well-being

One of the hypothesized explanations for productivity slowdown in US since the past decade is that existing economic indicators (including GDP) do not properly measure the advancements in information technologies. While average annual labor productivity growth was 2.8% per year over 1995-2004, it shrunk to 1.3% per year over 2005-2015 (Syverson 2016). An optimistic interpretation is that recent productivity gains due to innovations in IT-related goods and services are not properly reflected in the current productivity measures (e.g. Brynjolfsson and McAfee 2014, Aeppel 2015, Hatzius 2015). However, recent literature (Byrne et. al. 2016, Syverson 2016) has emphasized that while productivity mismeasurement may be important in recent years, it was also likely important in the past, so its power to explain the productivity slowdown is limited.

Our research does not aim to contribute directly towards this debate. Instead we focus on the more fundamental issue that GDP, and thus productivity, is not a direct measure of well-being in the first place. Thus, whether or not GDP or productivity mismeasurement has grown, is a distinct, albeit related, question from how well-being is changing. This has been an issue since GDP was invented and arguably an even bigger issue in the current digital era.

Perhaps no one has described the shortcomings of GDP[5] as a welfare measure as eloquently as Robert F. Kennedy:

> *Gross National Product counts air pollution and cigarette advertising, and ambulances to clear our highways of carnage. It counts special locks for our doors and the jails for the people who break them. It counts napalm and counts nuclear warheads and armored cars for the police to fight the riots in our cities...*
>
> *Yet the gross national product does not allow for the health of our children, the quality of their education or the joy of their play. It does not include the beauty of our poetry or the strength of our marriages, the intelligence of our public debate or the integrity of our public officials.*
>
> *It measures neither our wit nor our courage, neither our wisdom nor our learning, neither our compassion nor our devotion to our country, it measures everything in short, except that which makes life worthwhile.*[6]

Kennedy's poetic words themselves contribute much to our understanding (if not to our GDP!) and there have been a number of efforts to create a more comprehensive estimate of

---

[5] Kennedy was technically discussing GNP, but his comments are equally applicable to GDP.

[6] Robert Kennedy speaking at University of Kansas in 1968 (Ref: http://www.jfklibrary.org/Research/Research-Aids/Ready-Reference/RFK-Speeches/Remarks-of-Robert-F-Kennedy-at-the-University-of-Kansas-March-18-1968.aspx).

well-being. The country of Bhutan adopted a Gross National Happiness index since the 1980s aiming to measure the aggregate happiness in the country (Bates 2009). More recently, Jones and Klenow (2016) propose a measure that incorporates consumption, leisure, mortality and inequality to measure the economic well-being of a country.

In this paper, we are less ambitious and seek to stick more closely to a traditional microeconomic framework. In particular, we focus on the changes in consumer surplus generated by digital goods and discuss ways in which our approach can be expanded to more goods and services. Brynjolfsson and Saunders (2009) paraphrase Robert Solow in noting that the influence of the information age is seen everywhere except in the GDP statistics. Almost all of us use more and more digital goods such as search engines, smartphones, social networking sites, e-commerce platforms but their revenues don't always reflect this increased use. Consider the case of the music industry. Consumers shifted from buying physical units such as CDs, cassettes and vinyl records to downloading or streaming songs digitally through platforms such as iTunes, Pandora and Spotify. Digital goods have zero marginal cost and are hence priced much lower (often even zero price) than physical goods. Between 2004 and 2008 consumers listened to more music (units of music purchased increased from under 1 billion to over 1.5 billion without counting illegal downloads) but the recording industry's revenues declined by 40% (Brynjolfsson and Saunders 2009) and this trend has continued till now. Moreover, Waldfogel (2012) demonstrates that the quality of music has likely increased since 1999. Therefore, although the contribution of music industry to GDP statistics is shrinking consumers are better off than before; they are listening to more and better music.

The relationships among GDP, consumer surplus and well-being can be understood by looking at three illustrative cases. First, consider a situation that roughly describes many classic physical goods such as cars, consumer surplus is more or less proportional to firm revenue (Figure 1). Keeping the supply curve fixed, as the size of the market increases as more consumers enter the market, the demand curve simply shifts right. In this case, both consumer surplus and revenue increase roughly proportionately[7]. The increase quantity sold shows up in GDP statistics and hence both GDP and consumer welfare move in the same direction. Doubling the number of cars, apples or books sold is likely to roughly double revenues, GDP and consumer surplus. A similar logic applies for many services like haircuts, meals served or windows washed.

[Insert Figure 1 here]

A second case illustrates the situation faced by a number of traditional goods and services that are transitioning into digital goods and services. A good example of such a transition good is an encyclopedia. Previously browsing through a hard copy of an encyclopedia was often the best way to obtain information on a particular subject. Since 2000s, people have increasingly flocked to Wikipedia to get information about a wide variety of topics updated in real time by volunteers. In 2012, Encyclopedia Britannica, which had been one of the most popular encyclopedias, ceased printing books after 244 years (Pepitone 2012). Wikipedia has over 60 times as many articles as Britannica had, and its accuracy has been found to be on a par with Britannica (Giles, 2005). Far more people use Wikipedia than ever used Britannica. While the revenues from Britannica sales were counted in GDP statistics, Wikipedia has virtually no revenues and

[7] In the special case of horizontal supply curve and thus constant price, the effect is exactly proportional.

therefore doesn't contribute anything to GDP other than a few minimal costs for running servers and related activities and some voluntary contributions to cover these costs. Similarly, many people now have maps, GPS and other services available for no extra cost once they are able to access the Internet on mobile devices or home computers. For such transition goods, consumer surplus increases and revenue decreases as prices become zero (Figure 2). Hence GDP and consumer welfare move in opposite directions.

[Insert Figure 2 here]

A third category is the case of purely digital goods such as email, messaging apps, Facebook and Google search which have essentially zero marginal cost and are typically offered to the consumers for free. In some cases, digital goods earn revenues from advertising and this can create a significant contribution to GDP (Nakamura and Soloveichik 2015). However, changes in advertising revenues are generally not closely related to changes in consumer surplus (Spence and Owen 1977). Hence, as the demand for these free goods increases, consumer surplus will also increase but this change in well-being is not well-reflected in GDP (Figure 3). GDP may be completely unchanged due to this shift even though consumers are better off.

[Insert Figure 3 here]

More formally[8], consider the case of smartphones. Varian (2016) states that a smartphone is a substitute (to varying degrees) for a camera, GPS, landline, gaming console, ebook reader, personal computer, video and audio player, maps/ atlas, alarm clock, calculator, sound recorder etc. Consider the simplifying case of two goods available in two periods: a digital camera and a feature phone in period 1 and a digital camera and a smartphone in period 2. Suppose that the

[8] We thank Hal Varian for sharing his notes on GDP and Welfare which contained this example.

value of the camera to the consumer is v_1, the value of the feature phone is v_2 and the value of the smartphone is v_1+v_2. Assume that a device fully depreciates in a time period, i.e. a consumer has to purchase new devices each period. Also assume that a consumer buys both the camera and the feature phone in period 1 and only the smartphone in period 2 and there are a total of x such consumers. Suppose that the price of the camera is p_1 in period 1, the price of the feature phone is p_2 in period 1 and the price of the smartphone is also p_2 in period 2. Therefore, we have

(1) $(v_1-p_1)x + (v_2-p_2)x \geq 0$

(2) $(v_1+v_2-p_2)x \geq 0$

Subtracting (1) from (2) gives us p_1x which is the change in total consumer surplus from period 1 to period 2. This corresponds to the cost savings of not buying the digital camera since it is now included in the smartphone. However the contribution of these goods towards GDP (i.e. the market price of final goods) is $(p_1+p_2)x$ in period 1 and p_2x in period 2. Hence change in GDP from period 1 to period 2 is $-p_1x$ which is exactly the opposite of change in consumer surplus. Therefore while GDP goes down due to people not purchasing the digital camera, consumer surplus goes up.

In more concrete figures, the total number of digital cameras shipped worldwide dropped from 121 million units in 2010 to 24 million units in 2016[9]. In the same period, worldwide smartphone sales increased from 297 million in 2010[10] to 1.5 billion in 2016[11]. In the meanwhile, the total number of photos taken has also increased dramatically growing from 350 billion in

[9] http://www.cipa.jp/stats/dc_e.html
[10] http://www.gartner.com/newsroom/id/1543014
[11] http://www.gartner.com/newsroom/id/3609817

2010[12] to 2.5 trillion in 2016[13]. Moreover, the price for taking a photo with your smartphone is essentially 0 (compared to a positive price for printing photos in the analog era).

Figures 2 and 3, along with the illustrative example of smartphone, suggest that changes in consumer surplus is an important supplement to GDP as a measure of well-being for the current digital economy for either transition goods or purely digital goods. This is likely to become increasingly relevant as more and more goods transition from physical to digital in a variety of areas including financial advising (robo advisors such as Wealthfront), customer service (AI powered services such as DigitalGenius) and law (AI powered bots such as DoNotPay).

While producer surplus cannot be inferred from consumer surplus, when it comes to technological advances, firms are able to appropriate only a small fraction of the social returns (Nordhaus 2005) so we can focus on consumer surplus. If the share of producer surplus contribution to the total social surplus remains relatively stable, then our results would have to be scaled up only slightly if one wanted to estimate total surplus. However, Furman and Orszag (2015) provide evidence that the top performing companies have been earning increasingly larger returns to capital. Therefore, measuring simply changes in consumer surplus might underestimate changes in total surplus more significantly if the producer surplus grows relative to the consumer surplus.

[12] https://www.nytimes.com/2015/07/23/arts/international/photos-photos-everywhere.html

[13] https://www2.deloitte.com/global/en/pages/technology-media-and-telecommunications/articles/tmt-pred16-telecomm-photo-sharing-trillions-and-rising.html

2.2 Prior work measuring consumer surplus from digital goods

Recently there has been growing interest from researchers to estimate the changes in consumer surplus from digital goods. For instance, Greenstein and McDevitt (2011) estimate the additional consumer surplus created by broadband internet when consumers switched from dial-up to broadband. They estimate it to be between $4.8 and $6.7 billion from 1999-2006. For 2015, this figure is estimated to be $55 billion (Syverson 2016). Although this approach captures the welfare gains due to better internet access, it does not capture the increasing value of the digital information goods available online.

Another stream of literature has tried to measure the value of digital information goods by measuring the time spent using them. The underlying assumption behind these papers is that there is an opportunity cost associated with using free digital goods and this cost is equal to the wages lost due to not working. Therefore, the value of these digital goods is equal to these lost wages. Using this approach, Goolsbee and Klenow (2006) estimate the effect on consumer surplus for the median US resident to be $3000 for 2005. Brynjolfsson and Oh (2012) extend this method to include substitutability between online and offline goods (e.g. TV). After accounting for this, they estimate the average annual change in consumer surplus of the internet to be about $25 billion between 2007 and 2011. If consumers value their time less than their wages, then the consumer surplus may be overestimated.

Nakamura and Soloveichik (2015) estimate the value of free media by computing the online advertising revenues generated by websites. Including free media increases real GDP growth by 0.019% according to their estimates. However, advertising revenues do not capture the entire value of the free digital goods. For example, in 2011 Google earned around $36 billion ad revenue (Miller 2012) while Varian (2011) estimated the consumer surplus of Google to be between $65-$150 billion. Moreover, advertising revenues need not be proportional to consumer surplus. Spence and Owen (1977) argue that advertisers pay for number of views (or clicks) regardless of whether these views created low or high value for a consumer. For example, advertising revenues can be high for a program of broad interest (more views) but welfare need not be very high consumers might only be marginally interested in that program. However, for a niche program (less views), which is valued very highly by a small group of consumers, welfare will be high but advertising revenues will be low.

While these estimates of consumer surplus are based on available market data, our method will use choice experiments to elicit consumers' valuation of goods. Specifically, we will ask consumers to make a choice between keeping a digital good or taking a monetary compensation when foregoing it. This approach experimentally varies the monetary values and therefore addresses the limitation that the actual market price of many digital goods is zero so that demand does not represent their value. Moreover, an experimental setting may be better able to isolate consumers' valuation of goods compared to market data that is typically confounded by many other variables; albeit, depending on the design of the experiment, it may come at the expense of being "hypothetical", i.e., inconsequential (Carson and Groves 2007) and therefore either noisy or biased, as we discuss below.

3. Methodology

3.1 Approaches to measuring consumer value

There are two general approaches to obtain input data to measure consumer value: 1) based on market data ("revealed preferences") and 2) based on choice experiments or survey techniques ("stated preferences").

Approaches based on revealed preferences relate variation in observed demand for a good to changes in its market prices in order to derive demand curves and prices elasticity (e.g., Cohen et al., 2016; Greenwood and Kopecky 2011). Similarly, hedonic pricing models try to decompose the overall value of a good into the value contribution of its characteristics by applying regression-type models to the observed market prices and differences in characteristics of the goods (Williams 2008). However, both of these approaches require variance in the observed market prices and are therefore not applicable to digital goods that are provided for free. Revealed preference approaches are therefore limited to goods for which a market price exists, e.g., (broadband) internet access fees (Greenstein and McDevitt 2011) or require a proxy for market price, e.g., time spent using the digital goods (Goolsbee and Klenow 2006; Brynjolfsson and Oh 2012).

Stated preference elicitation techniques provide more flexibility because they do not require a market price or real transactions to exist and can be applied to contingent scenarios (leading to contingent valuation studies). Questions of stated preference include asking consumers directly about their maximum willingness-to-pay (WTP) in monetary terms. This question reveals a (potentially ratio-scaled) measure of a consumer's value of the good. However, this type of question has been shown to be less reliable and less valid, likely because consumers are not used to formulating own prices and do not have an incentive to reveal their true preferences (Miller et al. 2011; Carson and Groves 2007).

The introduction of non-hypothetical, incentive compatible variants to elicit WTP in form of auctions (e.g., Vickrey auctions, Vickrey, 1961) or lotteries (e.g., BDM, Becker, DeGroot, and Marschak 1964; Wertenbroch and Skiera 2002) mitigated some of these disadvantages, but at the expense of being complex, hard to understand, and by introducing (artificial) competitive pressure in auctions (Carson, Groves, and List 2014; Völckner 2006). These incentive compatible direct question formats are therefore not well-suited to either digital goods, in which supply is not restricted, or to large scale online choice experiments that consumers need to understand and answer quickly.

An alternative, indirect form of measuring stated preferences are discrete choice experiments (DCE) (Louviere, Hensher, and Swait 2000). DCEs ask consumers to choose between options and select the alternative that they value most. By experimental variation of the characteristics of the options (including prices) and applying logit or probit estimation models it is then possible to estimate consumers' utility function for the characteristics, i.e., their valuation of features and sensitivity to price changes. DCEs have become a synonym for choice-based conjoint experiments that typically involve about 8 to 12 sequential choice tasks that present multiple alternatives, e.g., three to five, and each alternative varies in multiple attributes (Rao 2014). These DCEs have a long tradition in, among others, marketing (e.g., value of product features), transportation (e.g., valuation of travel time savings), contingent valuation (e.g., valuation of preventing another Exxon Valdez type oil spill, Carson et al. 2003), and are also applied to economic valuation contexts (e.g., Rosston, Savage, Waldman 2011). They are widely relied upon in the legal proceedings to estimate values of goods for the purposes of damages calculations (e.g. in the Apple-Samsung lawsuit; see also McFadden 2014).

3.2 Proposed approach and implementation

We propose to measure consumer value of digital goods with DCEs. Instead of a conjoint-type experiment, we suggest a simpler implementation in which we only ask consumers to make a single choice among two options: Whether to keep access to a certain good or to give up the good and get paid a specific amount of money in return. We only ask one question per consumer and vary different price points systematically between consumers. The procedure can therefore be termed single binary discrete choice (SBDC) experiment (Carson and Groves 2007; Carson, Groves, and List 2014). We deliberately elicit only limited information from each consumer, i.e., data that is nominal-scaled, with the benefit that this information can be captured more reliably. Consumers only have to make a decision between two options instead of thinking about a monetary figure themselves. Moreover, we can compensate for the loss in information at the individual level by using large-scale choice experiments and aggregating the responses from the overall sample in order to derive ratio-scaled demand data. Thus we use large, and potentially massive, sample sizes to overcome some of the limitations of earlier research relying on smaller samples.

For the implementation, we use Google Consumer Surveys (GCS) as our primary platform. GCS allows us to run short one-question surveys cheaply and quickly and is therefore well suited for our SBDC experiments. A number of online publishers (including news and arts/entertainment sites) participate in GCS and host these choice experiments on their site as a gateway to access premium content (Stephens-Davidowitz and Varian 2015). Users have to answer the survey in order to unlock premium content (Figure 4). Survey creators pay per response, part of which goes to the publisher for hosting it. In addition to the responses, some demographic characteristics of the respondents such as region, age, gender and income are also provided which are inferred from IP address, location, browsing history (provided by Google's DoubleClick cookies which are also used to serve ads) and census data. Prior research has found that GCS results are very similar to those obtained from other surveys conducted by professional organizations such as Pew (Stephens-Davidowitz and Varian 2015).

[Insert Figure 4 here]

3.3 Utility theory and choice model

DCEs in general, including SBDC questions, are compatible with economic theory and can be used to estimate neoclassical Hicksian welfare measures (McFadden 1974, Carson and Czajkowski 2014). We will use utility theory and the random utility model to conceptualize the value that individual consumers obtain from consuming digital goods and the monetary value that they attach to them.

Specifically, we represent the utility that a consumer experiences from consuming a digital good g by $U(g)$. In our SBDC questions, utility is only affected by a change in the availability of the good with consumption quantities restricted to 1 and 0, i.e., a consumer can either use a good within a defined time period (g^1) or not (g^0). We abstract away from the intensity or duration of usage in this conceptual model but can account for it in our empirical application. We assume a constant market price of zero for the goods, which therefore does not have to be added to the utility function. We also do not explicitly consider the influence of other attributes such as negative utility effects of advertising or limited privacy (as analyzed in IAB Europe; McKinsey analysis) as they are nested within g^1. These components can be easily added to the utility function when they are subject to experimental variation. We further assume that $U(g^1) \geq U(g^0)$, i.e., that consumers derive a non-negative utility of consuming the good (and would otherwise not use it). A measure of monetary value can then be estimated by introducing two Hicksian measures, either the compensating measure, *C*, or the equivalent measure, *E*, that have an effect on the consumer's income y (Carson and Czajkowski 2014), such that:

(3) $U(g^1, y - C^*) = U(g^0, y)$, or

(4) $U(g^1, y) = U(g^0, y + E^*)$,

with C > 0 and E > 0.

C* is typically referred to as willingness-to-pay (WTP) for getting access to the good, while E* can be seen as willingness-to-accept (WTA) to forego it.

While, theoretically, C* should have the same magnitude as E*, empirical studies show that typically E* > C* (Hanemann 1991). It therefore becomes relevant to define the *status quo* of the valuation approach. When valuing the availability of free digital goods it seems reasonable to focus on WTA and assume that $U(g^1, y)$ is the *status quo* since using the good requires no upfront investment (y – C) from consumers.

When observing in the SBDC experiment that a consumer chooses to forego using a good for amount E instead of keeping it then we can assume that $U(g^0, y + E) > U(g^1, y)$, or $U(g^0, y +$

E) – $U(g^1, y) > 0$. Therefore, only differential effects need to be considered between the choice options so that the overall income can be excluded and only the marginal effect of E needs to be considered. Without loss of generality, we can then define the status quo utility as $U(g^1) = 0$. Consequently, a consumer will forego the good for amount E if $U(g^0, E)$ is positive, and will not if it is negative.

In order to estimate the equivalent measure E* we need estimates of how valuable consumers find using the good and how sensitive they react to changes in E. The random utility model is the standard framework to estimate the underlying utilities. It assumes that utility *U* consists of a systematic component *V* and a random component *e* that is inherent to consumer choice behavior and/or unobservable to the researcher (Manski 1977; Thurstone 1927), such that $U(g^0, E) = V(g^0, E) + e$. Typically, it is assumed that the systematic utility consists of part-worth utilities for each of the goods components, i.e., $V = b_0\, g^0 + b_1\, E$. The framework then allows to express the observed choices as probabilities *P* within a binary logit model, i.e., the probability that a consumer chooses to forego the service (or, on an aggregate level, the share of consumers who are willing to accept E) is:

(5) $P(g_0, E) = exp(b_0\, g^0 + b_1 E) / (1 + exp(b_0\, g^0 + b_1 E))$

or 1 - $P(g_0, E)$, for keeping the service. The parameters can then be estimated using closed-form maximum likelihood procedures. The median equivalent measure E* is then the price that makes consumers indifferent between the two options so that $P(g_0, E^*) = 0.5$ or $b_0\, g^0 + b_1 E = 0$, which leads to $E^* = - b_0\, g^0 / b_1$.

Here, we represent the utility function as linear in terms of monetary amounts. We will relax this assumption in the empirical application to handle non-linear terms and include further demographic variables.

3.4 Criticism

Such SBDC questions have several advantages compared to approaches that directly ask consumers about their WTP or WTA (i.e., C or E). SBDC valuations are based on consumer choices that are most similar to day-to-day (purchase or consumption) activities. They are natural manifestations of consumers' preferences and are easy to accomplish. DCEs have been shown to achieve good (external) predictive validity and produce valid estimates of WTP so that they should be favored over direct elicitations techniques (Carson and Groves 2007; Miller et al. 2011, Wlömert and Eggers 2016).

Moreover, SBDC questions are in line with economic theory and, based on the random utility model, can be used to estimate neoclassical Hicksian welfare measures (McFadden 1974, Carson and Czajkowski 2014). Moreover, a single "take-it-or-leave-it" referendum-like question has favorable incentive-compatibility properties, compared to multiple (e.g., "double-bounded"), sequential questions (Carson et al. 2003). Relatedly, single questions prevent the so-called starting point bias in follow-up questions, i.e., an anchoring effect in which subsequent prices are evaluated relatively to the first price the respondent was exposed to (Whitehead 2002).

However, the proposed SBDC questions, or rather contingent valuation questions in general, are not without criticism. Hausman (2012) identifies three major weaknesses of contingent valuation questions: 1) hypothetical bias; 2) differences between WTP and WTA; and

3) inconsistencies regarding scope and embedding (see a detailed rebuttal to his criticism from Haab et al. 2013). While the choice experiments we run in this paper differ somewhat from the contingent valuation approach that Hausman discusses, they have enough similarities that it is worth considering his critiques in some detail.

A hypothetical bias arises from SBDC questions (and contingent valuation questions) if consumers do not believe that they actually need to pay a price for any of the services in the near future. Hence, consumers might not consider the answer given to the stated preference question as consequential and have no incentive to reveal their true valuation so that a random response would be as a good as the true answer (Carson, Groves, and List 2014). According to the random utility model these answers increase the error variance so that preference estimates are less reliable. If these errors are unbiased around the true value, then they will tend to cancel out as sample size grows. However, apart from random error, it has been shown that the hypothetical bias often leads to a *systematic* bias, e.g., due to strategic incentives to understate or overstate the true WTP (Carson and Groves 2007). This is a more serious concern.

In order to quantify the magnitude of the hypothetical bias we test it empirically by providing consumers with real money if they stop using several services. In that way, choices are consequential and consumers have a clear incentive to provide their true valuation of the service (Carson, Groves, and List 2014). Similar applications of SBDC questions date back to a study by Bishop and Heberlein (1979) in which they measure the consumer surplus of goose hunting permits by providing hunters with actual money for returning their hunting permits. Similar approaches to mitigate the hypothetical bias include the incentive alignment procedure applicable to DCEs that has been shown to increase (external) validity substantially (Ding 2007;

Ding, Grewal, Liechty 2005; Wlömert and Eggers 2016). As discussed in our results section, we confirm that a hypothetical bias exists and that this criticism is justified. By measuring the hypothetical bias for several goods we present a rationale and approximation for a calibration factor in order to correct the bias.

While, theoretically, WTP and WTA should give (approximately) the same value, empirically they do not, which is recognized as being inconsistent with economic theory. Therefore, attempts have been made to extend (behavioral) theory in order to explain the disparities, e.g., with endowment effects, loss aversion, or uncertainty about the quality of the goods (Hanemann 1991; Plott and Zeiler 2005). Nevertheless, Hausman (2012) argues that the gap between WTP and WTA is "likely due to the reality that answers to contingent valuation surveys do not actually reflect stable or well-defined preferences but instead are opinions invented on the fly" (p. 47). This statement, however, is at odds with the fact that the differences between WTP and WTA are persistent and consistent, which rules out random effects of unstable preferences but rather results from systematic differences in these question formats. As we argue above, we find that WTA better represents consumer welfare since free digital goods require no upfront investment from consumers. Moreover, research from behavioral economics suggests that measuring WTP instead of WTA for free goods can lead to biased estimates since consumers may take the market price of zero as an informational "anchor" so that WTP estimates are also biased towards zero (Ariely, Loewenstein, and Prelec 2003).

The third major criticism, scope and embedding, refers to the proposition that consumers should be willing to pay more for a large effect than for a subset of that effect (or a good that is embedded in a larger package). Although these effects can be found empirically they are typically not considered large enough to be credible (Hausman 2012). Diamond and Hausman (1994) propose an adding-up test for the scope test. However, because digital goods can serve as substitutes or complements (e.g., social media can provide messaging functions or video can be used to listen to music) the adding-up test is not appropriate in our context. Differences in scope and embedding therefore need to be critically addressed in terms of the substitutability or complementarity of the goods. We address substitution effects in our empirical study (see section 4.5).

Much of the criticism refers to the nature of non-market goods that are typically subject of contingent valuation studies (e.g., clean air or clean water). With these goods, which may be used predominantly in a passive way, consumers have limited to no active experience or are rather unfamiliar with their true value (Carson 2012; Carson and Czajkowski 2014; Hausman 2012). Digital goods, on the other hand, are typically used on a day-to-day basis and required an active step in getting access to them (e.g., the deliberate choice to subscribe to Facebook) so that consumers should be familiar with them and are therefore better able to express and quantify their value. Morwitz et al. (2007) support this notion in their meta-analysis and empirical study. We therefore expect that our research context mitigates much of the above-mentioned criticism. However, arguably, it does not eliminate them completely. We will analyze the magnitude of the bias empirically.

4. Results

4.1 Google Consumer Surveys: Single Binary Discrete Choices

We identified the most widely used apps and websites on various devices and combined them into the following eight categories: Email, Search Engines, Maps, E-commerce, Video, Music, Social Media and Instant Messaging. We ran several SBDC surveys for each of these categories in 2016 where we asked consumers if they would prefer to keep access to the specific category or go without access to these goods for one year and get paid $E. We chose 10-15 price levels for each SBDC experiment and around 500 responses per price level. If the median WTA was outside the range of our price levels, we increased the number of price levels and ran additional choice experiments with different subjects.

We used binary logit models as outlined above for the estimation. Figure 5 plots the estimated WTA demand curves for these categories of digital goods. We show shares for keeping the goods instead of being willing to accept the money in order to be consistent to demand curves. That is, we plotted them in a way that makes it easier to see the negative effect of price. The demand curves appeared reasonable and fit the observed shares well. Table 1 provides the median WTA values per year. The available demographic variables (gender, age, income and urban density) reported by Google were also added to an extended model to determine variation in valuations of different groups. These extended logit models are reported in Table A.1 in the appendix.

According to our median *annual* WTA estimates, Search Engines ($16,629) seems to be the most valued category of digital goods followed by Email ($6,896) and digital Maps ($2,790). The way to interpret this is that 50% of the people in our sample would refuse to give up all access to search engines for one year even if we paid them about $16,000, while they would need to be paid a bit less than $7,000 in order to give up all access to email, and a bit less than $3000 for access to digital maps. One possible reason for this could be the lack of proper substitutes in the absence of search engines, email or digital maps compared to the other categories in our sample. Since most consumers do not directly pay for these services, almost all of the WTA for maps contributes towards consumer surplus. What's more, for many people, these services are essential to their jobs. Next, Video streaming services (e.g. Youtube, Netflix) are valued by consumers with a median WTA of $936 per year. Although, consumers do pay for some of these services, these amounts are of the order of $10-$20 per month (for those who pay) meaning the median consumer still enjoys an annual surplus of over $600. The remaining categories for which we estimated the median WTA are (in descending order) E-Commerce ($771), Social Media ($188), Instant Messaging ($145) and Music ($144).

All these estimates are potentially biased downwards due to lack of incentive compatibility as will be shown subsequently. When taken together, the sum of these estimates suggests there is a significant amount of consumer surplus from digital goods.

[Insert Figure 5 here]

[Insert Table 1 here]

Our approach can be used for digital and non-digital goods alike. As an example, we also ran SBDC surveys to estimate the WTA to give up the option of eating breakfast cereal[14] for 1 year. Figure 6 plots the WTA demand curve for breakfast cereal. We estimate the median WTA to give up breakfast cereal to be $48.46 per year in the US (95% CI: [$42.01, $55.60]). Examining a few non-digital goods can help us calibrate the relative importance of some of the digital goods we examine.

[Insert Figure 6 here]

4.2 Benchmark method: Best-worst scaling

As a benchmark to the GCS we conducted a choice experiment based on the best-worst scaling approach (Flynn et al. 2007; Marley and Louviere 2005). Best-worst scaling asks consumers to repeatedly select the best and worst options from sets of alternatives. Collecting more information, both within the choice set and across sequential choice sets, for each consumer makes this approach more efficient compared to the SBDC approach, which elicits only one decision. Moreover, consumers are required to make a tradeoff when deciding which goods they perceive as most and least valuable. This may mitigate or even eliminate the hypothetical bias, at least with respect to the ordinal ranking of the choices.

We used a list of 30 goods (a mixture of digital and non-digital goods), including six price points ranging from $10 to $10,000, that consumers had to evaluate. Since we examined the value of not having access to specific services or amenities for one year the prices were also

[14] Economist have studied this industry using a variety of approaches. See e.g. Hausman, 1996, Schmalensee, 1978, Nevo, 2001, and others.

formulated as losses, e.g., "earning $10,000 less for 1 year." The price sensitivity we are observing is therefore closer to WTP than WTA.

We used three options within each choice set so that respondents created a full ranking of the options by indicating the best and worst options. Figure 7 shows an exemplary choice set. Each respondent answered 10 sets in order to be exposed to each good. We randomized the allocation of goods and prices to choice sets across respondents.

[Insert Figure 7 here]

We recruited consumers for this online study via Peanut Labs, a professional panel provider with 2.9 million active panelists and member of, among others, CASRO, ESOMAR, and MRA (Peanut Labs 2015). We targeted consumers that are 18 years or older and live in the US. Consumers who did not fulfill these criteria were screened out. We controlled quotas for gender, age, and US regions to match US census data (File and Ryan 2014). In total, 204 respondents completed the study, leading to 4080 observed choices that were used in the estimation.

The estimation leads to interval-scaled utility scores that represent the relative value of the 30 goods (Figure 8, also Table A.2 in Appendix). We have set Facebook as a reference category so that utilities are expressed relative to Facebook. The estimated utility scores are strongly correlated with the WTA values derived from the GCS SBDC experiments (Correl. = 0.862). The ranking of the goods is also consistent between both approaches, with only one exception: online shopping is valued slightly more than video streaming in the best-worst scaling approach, while we find the opposite in the GCS surveys. When comparing the utilities of the services to the utility scores of the price levels we find, as expected, consistently lower implied

WTP values than WTA estimates according to the GCS survey. Overall, comparing the results of both approaches indicates convergent validity. We will provide additional evidence of convergent validity for the specific case of Facebook.

[Insert Figure 8 here]

4.3 Hypothetical bias, validation, and time sensitivity: The case of Facebook

We use Facebook as a case in order to address the magnitude of the hypothetical bias, benchmark the SBDC procedure against a BDM lottery approach, and analyze the effects of the time frame.

4.3.1 Hypothetical bias

In order to measure the hypothetical bias we applied a hypothetical scenario as in the Google Consumer Surveys and, in addition, used a second group of consumers that were asked the same question but with an additional incentive compatibility procedure. In order to make the SBDC question consequential for the consumer, we informed them that we will randomly pick one consumer out of every 200 respondents[15] and fulfill that person's selection. Specifically, we told respondents that if they choose "Keep access to Facebook" nothing will change for them, however, they will also not receive any money. If they choose "Give up Facebook and get paid $*E*", we promised them the money in cash provided that they do not access Facebook for one

[15] Carson, Groves, and List (2014) show that stochastically binding procedures (here: one out of every 200 respondents) do not significantly affect the results compared to deterministically binding procedures. We can confirm this result for our Facebook study in which we also tested a condition in which one out of every 50 respondents was selected (*E* was kept at $50 in this condition). We did not find significant differences in the choice behavior when varying the chances to win (Sig. = 0.236).

month. We further informed them about the procedure how we can monitor their Facebook online status remotely and the requirement to provide their email address (see Appendix for the exact question wording and monitoring process).

In order to keep the incentive compatibility procedure manageable for this paper we used a time frame of one month instead of one year as in the SBDC experiment. We varied $*E* across twelve[16] price points (*E* = 1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 1000).

We again recruited consumers for this study from Peanut Labs in 2016. We targeted consumers that are 18 years or older and live in the US. We further asked consumers to select all online services they have used in the last twelve months from a list of 14 options, including a non-existent online service. Consumers had to select Facebook in order to qualify for the survey; if they (also) selected the nonexistent service which we included in the survey they were disqualified. We have set quotas for gender, age, and US regions to match US census data (File and Ryan 2014) and applied post-stratification for education and household income.

Consumers who accessed the survey were randomly allocated to the incentive compatible (IC) or the non-incentive compatible (NIC) condition and, within each condition, to one of the tested price points. Allocation to the price points was also random, however we sampled the highest and lowest price points twice as often in order to obtain more reliable estimates for the endpoints of the demand function. We received 2997 complete responses (n_{NIC} = 1500, n_{IC} = 1497). The average completion rate was 93.3%: 96.1% for the NIC survey and 90.4% for the IC condition. The lower completion rate for the IC group is typical for surveys with incentive alignment and is likely due to consumers who are not willing to participate in it. However, in this

[16] In a follow-up study, we included additional price points, i.e., $0.01, $200, $500 and found consistent results.

case the differences to the NIC group are minor and, since we actively controlled quotas, they are not likely to produce a non-response bias.[17]

Figure 9 shows the share of consumers who prefer to keep using Facebook at the different price points, separated for both the IC and NIC condition. For very low prices, i.e., a price of $1, the IC and NIC condition produce almost identical shares, which is reasonable. For higher prices the disparities increase leading to consistently higher shares in the IC condition.

[Insert Figure 9 here]

In order to quantify the impact of the hypothetical bias we estimated a binary logit model that accounts for the magnitude of E (here, log(E) provided a better fit to the data), group membership (dummy variable for IC), and whether the IC group differs in sensitivity towards E. Table 2 shows the estimation results. The intercept represents the share of NIC consumers who prefer to keep Facebook at E = $1 (i.e., log(1) = 0). This share is estimated to be exp(1.178)/(1 + exp(1.178)) = 76.5%. The NIC group's utility decreases by 0.449 with every one-unit increase in log(*E*), leading to a median WTA_{NIC} = $13.80 per month. The IC consumers do not differ in the intercept (Sig. = 0.905) but they react significantly less sensitive towards differences in E. A one-unit increase in log(E) results in a utility decrease of only -0.449 + 0.140 = -0.309. Consequently, IC consumers are willing to accept a substantially larger amount: WTA_{IC} = $48.48

[17] Using Google Surveys as a basis for the massive online choice experiments might introduce a selection bias because selection criteria are less stringent compared to a professionally recruited sample based on representative quotas. To test this potential bias we compared the NIC group from the Peanut Labs sample to a Google Surveys sample (n = 1451). Because Google Surveys do not screen respondents if they are Facebook users or not, unlike in the Peanut Labs study, we matched the NIC group by accounting for the share of non-Facebook users. A binary logit model confirms that there are no significant differences between both samples, neither in terms of their intercept (Sig. = 0.991) nor sensitivity towards *E* (Sig. = 0.474). See appendix for details.

per month. Thus, the hypothetical WTA is understated in this research context and needs to be calibrated by a factor of 3.5.

[Insert Table 2 here]

We repeated the analysis in the lab of a large European university. The lab setting allowed us to vary more services. Specifically, we analyzed Facebook, Instagram, Snapchat, Skype, WhatsApp, LinkedIn, Twitter, and "all smartphone mapping apps." While it is possible to check the online status on Facebook remotely, checking usage of the other services required another approach. In order to induce incentive compatibility for these services, we informed respondents that they will have to change their password to a computer-generated code, which we keep secret from the respondent in a sealed envelope. For maps, we checked usage statistics of the apps on the respondents' smartphones (given their permission). We inferred that the respondents have not used the services if the seal on the envelope was still intact and the password still active (not reset) and/or the services were not used according to the apps usage statistics. In total, 537 students completed the study, 66% of them in an incentive compatible setting for a 1 out of 50 chance to be selected and have the SBDC answer for one randomly determined service fulfilled.

The estimation results are consistent to the results for Facebook reported above (see Table 3). We find that none of the IC conditions differ significantly from the NIC condition in the intercept. Only the price sensitivity is affected. Students were consistently less price sensitive in the IC condition, however not significantly for Skype (Sig. = 0.780), maps (Sig. = 0.824), LinkedIn (Sig. = 0.232), and Twitter (Sig. = 0.088).

Table 4 shows the inferred WTA in the NIC and IC conditions. WTA for Facebook is higher in the lab study compared to the previous study, i.e., €106 per month in the IC setting, likely due to the different sample characteristics. The highest WTA (€709 per month) is achieved by WhatsApp, which 99% of the participants are using. The magnitude of the hypothetical bias calibration factor for these two services is around 4 and comparable to the previous study. However, for the remaining services we find lower calibration factors closer to or even below 1.0.

These calibration factors are strongly correlated with the intercepts of the models. The intercept is linked to the choice probability of giving up the service for a price of €1 and therefore indicates the general attractiveness of the service. An approximation of the calibration factor of the magnitude $exp(0.85 * \text{intercept} - 0.12 * \text{intercept}^2)$ is able to explain 95% of the hypothetical bias in WTA calculations, however, excluding the smartphone mapping apps. Maps deviate from this model. Potential reasons could be that it was the only service that was not specified further, e.g., by a certain brand name. Maps are also not password-protected so that the incentive alignment relied on checking usage statistics of the apps on the smartphone alone. However, maps can typically also be accessed through the browser, which would not show in the usage statistics. Taken together, the incentive compatibility of the maps scenario might have been less convincing than for the other services.

The calibration is applicable to non-incentive compatible models since the intercepts do not differ between NIC and IC scenarios. E-Commerce, as an example, which is worth $771 according to the GCS survey would be valued at $2378 when accounting for the hypothetical bias. We will return to the hypothetical bias and a rationale for the magnitude of the calibration factor in the sensitivity analysis section.

[Insert Table 3 here]

[Insert Table 4 here]

4.3.2 Benchmark method: BDM lottery

As another benchmark to the SBDC approach, we applied an incentive compatible BDM lottery procedure (Becker, DeGroot, and Marschak 1964) in order to elicit direct, numeric responses from consumers about their WTA. Specifically, we asked consumers about the minimum amount of money they would request in order to give up Facebook for 1 month. In order to achieve incentive compatibility we informed respondents that the amount will serve as their bid in a lottery. The BDM lottery instructs that, after the survey, a random price will be drawn from a uniform distribution of values. If the random price is higher than the bid, the respondent will be paid the random price when giving up Facebook for 1 month. If the random price is lower than the bid, the respondent will receive no money but can keep the access to Facebook.

We conducted the BDM lottery in the lab of the European university in parallel to the SBDC experiments. In total, 139 students took part in the lottery. We compare this sample to the respondents who took part in the incentive compatible SBDC experiment of the lab (n = 356). Figure 10 shows the demand functions that result from both approaches. The BDM lottery leads to a more detailed depiction of the demand function, including price thresholds. The SBDC derived function is closely aligned. The observed shares correlate strongly (Correl. = 0.891) and are not significantly different from each other (Sig. = 0.081).

[Insert Figure 10 here]

4.3.3 Effect of the analyzed time frame

In the incentive compatible studies we used one month instead of one year as the time frame that respondents should forego Facebook. This raises the question to what extent consumers are sensitive to time. To address this question, we conducted SBDC experiments in an incentive compatible setting in which, in addition to prices E, we varied the time frame across three periods, i.e., T = 1 week, 2 weeks, 1 month. We sampled respondents from Peanut Labs in 2017 using the same criteria as in the previous studies, however, we did not screen out respondents who do not use Facebook (assuming that these respondents have a low valuation for Facebook). A total of 1499 respondents were available for the analysis.

Table 5 shows the estimation results. As expected, the time frame has a significant positive effect on the probability to keep Facebook. Accordingly, the median WTAs for the different time frames are \$3.92 for one week, \$10.53 for two weeks, and \$17.61 for one month. These values and the beta estimates suggest that the effect of time is not necessarily linear. In order to get a better overview of the effect of time we sampled 5021 additional respondents via GCS in a non-incentive compatible setting. We allocated these respondents randomly to one of ten conditions that differ in the time frame: T = 1 hour, 1 day, 1 week, 2 weeks, 3 weeks, 4 weeks, 1 month, 2 months, 3 months, 6 months, 1 year. We kept E constant at \$50. Figure 11 shows the observed shares of respondents who prefer to keep Facebook at the different time frames and the estimated time function according to the binary logit model (using log(T)). It confirms the non-linear effect of time and that its positive marginal effect decreases with longer time frames, consistent with temporal discounting models.

[Insert Table 5 here]

[Insert Figure 11 here]

4.4 Sensitivity analysis

4.4.1 Effect of random answers

Random answers increase the error variance in choice model estimations. The error variance, in turn, has a negative effect on the precision, i.e., scale of the estimates *S* in logit choice models (Hauser, Eggers, Selove 2016). Specifically, the scale *S* is inversely proportional to the error variance. The scale *S* cannot be separately identified, such that it is incorporated in the estimated betas:

$V = (S * b_0)\, g^0 + (S * b_1)\, E.$

Lower scaled estimates (more error), i.e., estimates with lower magnitude, cause the logit function to become more linear. Higher scaled estimates (less error) lead to a stepwise function that allows to predict decisions and identify the the median WTA more precisely (see Figure 12).

[Insert Figure 12 here]

The effect is demonstrated in Table 6. The table shows the result of a modified bootstrapping procedure in which subsamples were drawn from the IC Facebook sample. In each subsample we replaced *R* random original responses with the same amount of *random* answers and estimated the logit model. This process was repeated 1000 times in order to be able to observe the variance of the results. The results show that more random answers decrease the scale of the estimates. The scale *S* is proportional to the relative share of non-random answers. Having more random answers than original responses (R = 800) causes the magnitude of the estimates to be less than half the size of the original estimation without additional random answers (R = 0). However, the mean WTA as well as the absolute standard error of the estimates remain largely unaffected. The same results are obtained for the NIC group.

[Insert table 6 here]

To get an estimate about the magnitude of error in the overall sample, we compared how many choices can be recovered correctly based on the estimated probabilities and a cutoff value of 0.5. Overall this measure results in the internal hit rate of correct predictions. For the IC group the hit rate is 62.1%, i.e., 37.8% of the choices cannot be predicted correctly based on the amount of E alone. Interestingly, the hit rate for the NIC group is higher, 69.9%. This suggests that consumers in the IC group faced a decision that was more difficult to make, likely because their choices were consequential. It is important to note that the misclassified choices are not necessarily due to purely random responses. These cases can also be explained by heterogeneity among consumers, either with respect to their valuation of Facebook or regarding their general price sensitivity (or both), which is not accounted for in the estimation models.

We find limited evidence that respondents answer purely at random. Using GCS (n = 502), we asked a question in which we requested respondents to select all services that they have used in the last 12 months. Only 1% of the respondents have chosen a non-existing service, i.e., answered randomly.

Figure 12 can also serve as an explanation for the calibration factors we calculated for the hypothetical bias. It can be assumed that in a hypothetical setting respondents are more likely to answer randomly, e.g., according to $S = 0.5$. Using an incentive compatible setting reduces random answers because choices are consequential so that the precision increases, e.g., to $S = 1.0$. When the valued good is attractive and has a utility of $V > 0$ the hypothetical scenario underestimates the true choice probabilities. The results therefore need to be calibrated upwards. This effect reverses when the good is less attractive, i.e., $V < 0$. In this case, the hypothetical scenario overestimates the true value so that calibration factors should be smaller than 1. This rationale is consistent with our empirical findings and would also explain the large variance in calibration factors given in the meta analyses of hypothetical biases by Murphy et al. (2005) and List and Gallet (2001).

4.4.2 Effect of sample size

The reliability of the WTA estimates depends on the sample size. To analyze the magnitude of the effect we used bootstrapping with varying subsample sizes to observe the effect on standard errors and confidence intervals for the WTA estimate. Each subsample of a given size was again randomly drawn 1000 times from the original sample. Table 7 demonstrates that the standard errors of the estimates are reduced by the square-root of 2 when doubling the sample size (in this case the scale of the estimates remains largely unaffected). This general pattern also holds for the standard error of the WTA estimate. However, since WTA is a ratio of two stochastic variables this generalization is approximate. The results show how the 95% confidence interval narrows when increasing the sample size. There is uncertainty in the measure even with a sample size of 1500. A 95% confidence interval of ±$10 would be achieved with a sample of 6000 consumers.

[Insert Table 7 here]

4.5 Facebook consumer valuation and growth calculation

Based on the incentive compatible Facebook survey (fielded with Peanut Labs in 2017, n = 1499), we added usage and demographic variables to further understand differences in consumer value. The estimation results can be found in Table 8. The number of minutes that consumers spend on Facebook per day (on a log scale, capped at 120 minutes) is a strong predictor for the value of Facebook. The more minutes respondents spend on Facebook the more likely they are to keep their access. There is a negative effect for the overall time that the respondent is registered on Facebook (measured on a 7-point scale from “less than 1 year” to “more than 10 years”) so that users who registered more recently (a longer time ago) are more (less) likely to keep Facebook. Interestingly, we do not find a significant effect for the number of friends someone has on Facebook (measured on a 6-point scale from “less than 50” to “more than 1000”). We also controlled for substitution effects due to other social media services, i.e., Instagram, Snapchat, and Twitter. The results confirm a significant substitution effect only for Instagram, such that respondents who have an Instagram account are more likely to give up Facebook.

In terms of socio-demographics, we find significant effects for gender and age of the respondent, as well as household income. Education and US region are not significant (not shown in Table 8). Specifically, we see that female respondents are more likely to keep Facebook. The same holds for older consumers. Household income also appears to have a positive effect, however the pattern is less consistent and the effect for the highest income category is only marginally significantly different from the lowest category (Sig. = 0.094). The effect is significant for consumers who preferred not to disclose their income.

[Insert Table 8 here]

We can use the number of minutes of Facebook usage per day as a proxy to sketch consumer valuation over time. Average Facebook usage has increased from 31 minutes per day in 2012 to over 50 minutes per day in 2016 (see Figure 13).[18] Calculating the change in WTA due to increased intensity of usage (*ceteris paribus*) suggests that consumers' valuation of Facebook increased from a median of $41.44 per month per user in 2012 ($497 per year) to $63.28 per month in 2016 ($759 per year, see Table 9). Multiplying these figures with the number of active users in the US, we get an annual increase in the value of Facebook of $18 billion from 2015 to 2016.

[Insert Figure 13 here]

[Insert Table 9 here]

5. Discussion

With advances in information technologies, we can now gather data at a large scale in close to real time. Initiatives such as MIT's Billion Prices project[19] and Adobe's Digital Price Index[20] are collecting price data from online retailers in real time to compute price and inflation

[18] http://www.statista.com/statistics/256343/monthly-minutes-spent-on-facebook-by-us-users-by-channel/, http://www.bloomberg.com/news/articles/2014-07-23/facebook-posts-second-quarter-revenue-profit-topping-estimates, http://www.nytimes.com/2016/05/06/business/facebook-bends-the-rules-of-audience-engagement-to-its-advantage.html?_r=0, http://www.nu.nl/files/IDC-Facebook%20Always%20Connected%20(1).pdf.
[19] http://bpp.mit.edu/
[20] https://blogs.adobe.com/digitalmarketing/analytics/introducing-digital-economy-project/

indices. We explore the potential to reinvent and supplement the measurement of economic well-being by taking advantage of the ease of gathering data in the digital era. The end goal of this research agenda is to design a scalable method of measuring changes in consumer surplus due to technological advancements. We explore a potential way of measuring changes in consumer surplus through SBDC experiments. Our method is highly scalable and relatively inexpensive. Therefore, it can be run at very frequent, regular intervals to keep track of changes in consumer surplus. As argued previously, this measure can be an important complementary indicator of consumer well-being for the digital economy.

In a series of online experiments we show that the SBDC approach leads to reasonable demand functions that are consistent to other validated approaches. Overall, we find that (free) digital goods provide substantial value to consumers. We further show that the approach can detect consumers' sensitivity towards different time frames. We show that time has a positive effect on the probability to keep a service with decreasing marginal returns. The decreasing effects are consistent with temporal discounting models. Nevertheless, consumers are increasingly willingly to undergo "digital detox" for a short duration by giving up internet or individual services like Facebook either through self-control or by installing software which blocks particular sites. This itself raises interesting questions about rationality and the nature of utility functions. Due to this trend we recommend to use longer time frames for the evaluation, e.g., one year, which would also lead to a more conservative estimate of WTA.

In order to address the limitation of a hypothetical bias of the proposed approach we have tested several goods in an incentive compatible setting. We confirm that a hypothetical bias exists and provide an approximation for a calibration factor. The calibration can explain the differences in the valuation between incentive compatible and non-incentivized settings well. However, the generalizability of such a correction needs to be analyzed further in future studies.

A major limitation of our study remains the lack of precision in our estimates. While the BEA is able to measure GDP very precisely (e.g. US GDP was reported as $16,514,593,000 on the first day of 2016), we are only able to provide a relatively coarse estimate of changes in consumer surplus. We therefore require large sample sizes to narrow the confidence interval of the WTA estimates. Future applications could explore adaptive approaches that adjust the analyzed price intervals dynamically in order to find the relevant price range for finding the median WTA and therewith save sampling costs. That said, our approach is at least attempting to directly measure a concept that is not correctly measured by other official data. In short, we believe it is better to be approximately correct than precisely wrong. Another limitation of our study is that it is biased towards people using the internet. Our choice experiments are only accessible online, therefore people not using the internet at all are excluded. Pew estimates that about 15% of Americans don't use the internet.[21] Accordingly, our results must be interpreted as relevant to this audience, but not necessarily others.

[21] http://www.pewresearch.org/fact-tank/2015/07/28/15-of-americans-dont-use-the-internet-who-are-they/

References

Aeppel, T. (2015). Silicon Valley Doesn’t Believe U.S. Productivity Is Down. The Wall Street Journal. (Accessed at: http://www.wsj.com/articles/silicon-valley-doesnt-believe-u-s-productivity-is-down-1437100700)

Ahlfeldt, G., P. Koutroumpis, T. Valletti, (2016) Speed 2.0: Valuing universal access to digital highways, Journal of the European Economic Association.

Ariely, Dan; Loewenstein, George and Prelec, Drazen. “ ‘Coherent Arbitrariness’: Sta- ble Demand Curves without Stable Preferences.” Quarterly Journal of Economics, February 2003, 118 (1), pp. 73-105.

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

Bates, W. (2009). Gross national happiness. Asian−Pacific Economic Literature, 23(2), 1-16.

Becker, G.M., DeGroot, M.H., & Marschak, J. (1964). Measuring Utility by a Single-Response Sequential Method. Behavioral Science, 9, 226–232.

Bishop, R.C. and Heberlein, T.A., 1979. Measuring values of extramarket goods: Are indirect measures biased?. American journal of agricultural economics, 61(5), pp.926-930.

Brynjolfsson, E. and Oh, J., 2012. The attention economy: measuring the value of free digital services on the Internet.

Brynjolfsson, E., & McAfee, A. (2014). The second machine age: Work, progress, and prosperity in a time of brilliant technologies. WW Norton & Company.

Brynjolfsson, E., & Saunders, A. (2009). What the GDP Gets Wrong.(Why Managers Should Care). MIT Sloan Management Review, 51(1), 95.

Carson, R. and Czajkowski, M., 2014. The discrete choice experiment approach to environmental contingent valuation, in: Hess, S. and Daly, A. eds., 2014. Handbook of choice modelling. Edward Elgar Publishing.

Carson, R.T., 2012. Contingent valuation: A practical alternative when prices aren't available. The Journal of Economic Perspectives, 26(4), pp.27-42.

Carson, R.T., Groves, T., 2007. Incentive and informational properties of preference questions, Environmental and Resource Economics, 37(1), pp.181-210.

Carson, R.T., Groves, T., List, J.A., 2014. Consequentiality: A theoretical and experimental exploration of a single binary choice, Journal of the Association of Environmental and Resource Economists, 1(1/2), pp.171-207.

Carson, R.T., Mitchell, R.C., Hanemann, M., Kopp, R.J., Presser, S. and Ruud, P.A., 2003. Contingent valuation and lost passive use: damages from the Exxon Valdez oil spill. Environmental and resource economics, 25(3), pp.257-286.

Cavallo, A., & Rigobon, R. (2016). The Billion Prices Project: Using online prices for measurement and research. The Journal of Economic Perspectives, 30(2), 151-178.

Cohen, P., Hahn, R., Hall, J., Levitt, S., & Metcalfe, R. (2016). Using Big Data to Estimate Consumer Surplus: The Case of Uber (No. w22627). National Bureau of Economic Research.

Diamond, P.A. and Hausman, J.A., 1994. Contingent valuation: Is some number better than no number?. The Journal of economic perspectives, 8(4), pp.45-64.

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

Ding, M., Grewal, R. and Liechty, J., 2005. Incentive-aligned conjoint analysis. Journal of marketing research, 42(1), pp.67-82.

File, T. and Ryan, C. 2014. Computer and Internet Use in the United States: 2013, U.S. CENSUS BUREAU, http://www.census.gov/history/pdf/2013computeruse.pdf

Flynn, T. N., Louviere, J. J., Peters, T. J., & Coast, J. (2007). Best–worst scaling: what it can do for health care research and how to do it. Journal of Health Economics, 26(1), 171-189.

Furman, J., & Orszag, P. (2015). A Firm-Level Perspective on the Role of Rents in the Rise in Inequality.

Giles, J. (2005). "Internet encyclopaedias go head to head: Jimmy Wales' Wikipedia comes close to Britannica in terms of the accuracy of its science entries". Nature 438 (7070): 900–1

Goolsbee, A. and Klenow, P.J., 2006. Valuing consumer products by the time spent using them: An application to the Internet (No. w11995). National Bureau of Economic Research.

Greenstein, S. and McDevitt, R.C., 2011. The broadband bonus: Estimating broadband Internet's economic value. Telecommunications Policy, 35(7), pp.617-632.

Greenstein, S., & McDevitt, R. C. (2009). The broadband bonus: Accounting for broadband Internet's impact on US GDP (No. w14758). National Bureau of Economic Research. Chicago

Greenwood, J. and Kopecky, K.A. 2011. “Measuring the Welfare Gain from Personal Computers: A Macroeconomic Approach” Federal Reserve Bank of Atlanta Working Paper No. 2011-05. Available at SSRN: http://ssrn.com/abstract=1800266

Haab, T.C., Interis, M.G., Petrolia, D.R. and Whitehead, J.C., 2013. From hopeless to curious? Thoughts on Hausman's “dubious to hopeless” critique of contingent valuation. Applied Economic Perspectives and Policy, p.ppt029.

Hanemann, W.M., 1991. Willingness to pay and willingness to accept: how much can they differ?. The American Economic Review, 81(3), pp.635-647.

Hatzius, J. (2015). Productivity paradox 2.0. Goldman Sachs Global Macro Research, 39, 6-7. (Accessed at: http://www.goldmansachs.com/our-thinking/pages/macroeconomic-insights-folder/the-productivity-paradox/report.pdf)

Hauser, J.R., Eggers, F., and Selove, M. (2016), The Strategic Implications of Precision in Conjoint Analysis, Cambridge, MA: MIT Sloan School of Management

Hausman, J., 2012. Contingent valuation: from dubious to hopeless. The Journal of Economic Perspectives, 26(4), pp.43-56.
Jones, C. I., & Klenow, P. J. (2016). Beyond GDP? Welfare across countries and time. The American Economic Review, 106(9), 2426-2457.
Kuznets, S. (1934) “National Income, 1929-1932” 73rd Congress, 2nd Session, Senate Document no. 124.
Kuznets, S. (1973). Modern economic growth: findings and reflections. The American economic review, 63(3), 247-258.
Landefeld, J. S. (2000). GDP: One of the Great Inventions of the 20th Century. Survey of Current Business, 80(1), 6-14.
List, J.A. and Gallet, C.A., 2001. What experimental protocol influence disparities between actual and hypothetical stated values?. Environmental and Resource Economics, 20(3), pp.241-254.
Louviere, J.J., Hensher, D.A. and Swait, J.D., 2000. Stated choice methods: analysis and applications. Cambridge University Press.
Manski, C.F., 1977. The structure of random utility models. Theory and decision, 8(3), pp.229-254.
Marley, A. A. J. and Louviere, J. J., 2005. Some probabilistic models of Best, Worst, and Best-Worst choices. Journal of Mathematical Psychology. 49, 464-480.
McFadden, D.L. (1974), ‘Conditional logit analysis of qualitative choice behavior’, in P. Zarembka (ed.), Frontiers in Econometrics, New York: Academic Press.
McFadden, D.L. (2014), In the Matter of Determination of Rates and Terms for Digital Performance in Sound Recordings and Ephemeral Recordings (WEB IV), Before the Copyright Royalty Board Library of Congress, Washington DC, Docket No. 14-CRB-0001-WR, October 6.
Miller, K.M., Hofstetter, R., Krohmer, H. and 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), pp.172-184.
Miller, M. (2012). How Google Made $37.9 Billion in 2011. Search Engine Watch. (Accessed at: https://searchenginewatch.com/sew/news/2140712/google-usd379-billion-2011)
Molnar, G., Savage, S., & Sicker, D. (2015). Reevaluating the Broadband Bonus: Evidence from Neighborhood Access to Fiber and United States Housing Prices. Available at SSRN 2628396.
Morwitz, V.G., Steckel, J.H. and Gupta, A., 2007. When do purchase intentions predict sales?. International Journal of Forecasting, 23(3), pp.347-364.
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), pp.313-325.
Nakamura LI, and Soloveichik R (2015) Valuing ‘free’ media across countries in GDP. Federal Reserve Board of Philadelphia Working Paper No. 15-25.
Nordhaus, W. D. (2005). Schumpeterian Profits and the Alchemist Fallacy.Yale Economic Applications and Policy Discussion Paper, (6).

OECD (2008). OECD compendium of productivity indicators. (Accessed at: http://www.oecd.org/std/productivity-stats/40605524.pdf)

Peanut Labs (2015). Peanut Labs Panel Book, http://news.peanutlabs.com/rs/peanutlabs/images/PL-PanelBook_2015.pdf

Pepitone, J. 2012. Encyclopedia Britannica to stop printing books. CNN. Available at: http://money.cnn.com/2012/03/13/technology/encyclopedia-britannica-books/ [Accessed March 21, 2016].

Plott, C.R. and Zeiler, K., 2005. The Willingness to Pay–Willingness to Accept Gap, the. The American Economic Review, 95(3), pp.530-545.

Rao, V.R., 2014. Applied conjoint analysis (p. 56). New York, NY: Springer.

Rosston, G., Savage, S. and Waldman, D., 2011. Household demand for broadband internet service. Communications of the ACM, 54(2), pp.29-31.

Spence, M., B. Owen. 1977. Television programming, monopolistic competition, and welfare. Q. J. Econ. 103–126.

Stephens-Davidowitz, S., H. Varian. 2015. A Hands-on Guide to Google Data. Available at: http://people.ischool.berkeley.edu/~hal/Papers/2015/primer.pdf [Accessed March 21, 2016].

Stiglitz, J. E., Sen, A., & Fitoussi, J. P. (2010). Report by the commission on the measurement of economic performance and social progress. Paris: Commission on the Measurement of Economic Performance and Social Progress.

Syverson, C. (2016). Challenges to Mismeasurement Explanations for the US Productivity Slowdown (No. w21974). National Bureau of Economic Research.

Thurstone, L.L., 1927. A law of comparative judgment. Psychological review, 34(4), p.273.

Varian H (2011, March 29) The economic value of Google. Presentation, San Francisco.

Varian, H. (Sep 2016). A microeconomist looks at productivity: A view from the valley. Presentation, Brookings (Accessed at: https://www.brookings.edu/wp-content/uploads/2016/08/varian.pdf)

Vickrey, W. (1961). Counterspeculation, Auctions and Competitive Sealed Tenders. Journal of Finance, 16, 8–37.

Völckner, F., 2006. An empirical comparison of methods for measuring consumers' willingness to pay. Marketing Letters, 17(2), pp.137-149.

Waldfogel, J. (2012). Copyright Protection, Technological Change, and the Quality of New Products: Evidence from Recorded Music since Napster.Journal of Law and Economics, 55(4), 715-740.

Wertenbroch, K., & Skiera, B. (2002). Measuring Consumers' Willingness to Pay at the Point of Purchase. Journal of Marketing Research, 39, 228–241.

Whitehead, J.C., 2002. Incentive incompatibility and starting-point bias in iterative valuation questions. Land Economics, 78(2), pp.285-297.

Williams, B. 2008. " A Hedonic Model for Internet Access Service in the Consumer Price Index." Monthly Labor Review, July, pp.33-48.

Wlömert, N. and 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), pp.195-210.

Tables and Figures

Table 1: Median WTA Estimates

Category	WTA/ year	95% CI lower	95% CI upper	n
All Search Engines	$16,628.97	$12,469.69	$22,561.61	8081
All Email	$6,895.80	$5,602.04	$8,508.91	8097
All Maps	$2,789.88	$2,008.42	$3,906.18	6526
All Video	$935.99	$757.54	$1,165.04	6572
All E-Commerce	$770.91	$621.50	$973.99	6530
All Social Media	$187.79	$146.62	$238.84	6556
All Messaging	$144.75	$107.12	$193.47	6600
All Music	$144.16	$116.87	$179.07	6527

Table 2: Estimation results of binary logit model comparing IC and NIC group

	beta	Std. Error	z	Sig.
(Intercept)	1.178	0.135	8.726	<0.001
log(E)	-0.449	0.034	-13.147	<0.001
IC	0.022	0.184	0.119	0.905
IC*log(E)	0.140	0.045	3.076	0.002

Table 3: Lab study estimation results

	Facebook	Instagram	Snapchat	Skype	Whatsapp	Maps	LinkedIn	Twitter
(Intercept)	3.159	0.771	0.304	-0.440	4.936	2.593	0.380	-1.363
log(E)	-0.971	-0.590	-0.520	-0.493	-0.963	-0.688	-0.652	-0.587
IC*log(E)	0.293	0.178	0.151	0.025	0.212	0.011	0.096	0.245

Table 4: Hypothetical bias in the lab study

	Facebook	Instagram	Snapchat	Skype	Whatsapp	Maps	LinkedIn	Twitter
WTA NIC	€25.91	€3.70	€1.79	€0.41	€167.81	€43.37	€1.79	€0.10
WTA IC	€105.79	€6.50	€2.28	€0.39	€709.19	€46.03	€1.98	€0.02
Calibration factor	4.084	1.758	1.271	0.954	4.226	1.061	1.105	0.189

Table 5: Estimation results for the marginal effect of time

	beta	Std. Error	z	Sig.
(Intercept)	0.324	0.126	2.572	0.010
log(E)	-0.237	0.024	-10.009	<0.001
Time 1 week (reference)	(0.000)			
Time 2 weeks	0.235	0.135	1.734	0.083
Time 1 month	0.357	0.133	2.688	0.007

Table 6: Effect of random answers

Random sample R	Non-random sample	**Mean intercept**	**Mean beta log (E)**	Std. error Intercept	Std. error beta log(E)	Mean WTA	**Scale S**
800	700	**0.517**	**-0.135**	0.139	0.033	$46.29	**0.431**
400	1100	**0.846**	**-0.218**	0.149	0.035	$48.52	**0.700**
200	1300	**1.020**	**-0.262**	0.151	0.037	$49.06	**0.844**
100	1400	**1.122**	**-0.289**	0.157	0.038	$48.91	**0.929**
0	1500	**1.206**	**-0.311**	0.163	0.039	$48.18	**(1.000)**

Table 7: Effect of sample size

Sample size	Mean intercept	Mean beta log (E)	**Std. error Intercept**	**Std. error beta log(E)**	mean WTA	**95% CI lower**	**95% CI upper**
200	1.242	-0.319	**0.462**	**0.110**	$49.65	**$13.13**	**$187.73**
400	1.227	-0.316	**0.324**	**0.077**	$48.72	**$21.16**	**$112.28**
800	1.214	-0.311	**0.226**	**0.053**	$49.30	**$27.83**	**$87.27**
1500	1.206	-0.311	**0.163**	**0.039**	$48.18	**$31.69**	**$73.26**

Table 8: Facebook value diagnostic

	beta	Std. Error	z	Sig.
(Intercept)	0.179	0.378	0.473	0.636
log(E)	-0.295	0.030	-9.861	<0.001
Time 1 week (reference)	(0.000)			
Time 2 weeks	0.237	0.154	1.543	0.123
Time 1 month	0.390	0.152	2.575	0.010
Facebook usage minutes per day (log)	0.229	0.051	4.446	<0.001
Facebook number of friends (scale)	0.084	0.053	1.593	0.111
Facebook time registered (scale)	-0.126	0.048	-2.639	0.008
User of Instagram	-0.366	0.157	-2.330	0.020
User of Snapchat	-0.203	0.170	-1.194	0.232
User of Twitter	0.025	0.138	0.178	0.859
Gender female (reference)	(0.000)			
Gender male	-0.390	0.127	-3.072	0.002
Age 18-24 (reference level)	(0.000)			
Age 25-34	0.014	0.248	0.055	0.956
Age 35-44	0.297	0.240	1.238	0.216
Age 45-54	0.481	0.240	2.005	0.045
Age 55-64	0.438	0.253	1.734	0.083
Age 65+	1.074	0.278	3.868	<0.001
Income less than 25K (reference)	(0.000)			
Income 25K to 50K	0.220	0.220	1.000	0.318
Income 50K to 100K	0.052	0.213	0.244	0.807
Income 100K to 150K	0.261	0.245	1.064	0.287
Income 150K or more	0.444	0.265	1.674	0.094
Income “prefer not to answer”	0.811	0.388	2.091	0.037

Table 9: Annual change in value of Facebook

Year	Number of active US users (in millions)	Median WTA per user per year	Annual Change in Value (in millions)
2012	183	$497.30	
2013	195	$587.76	$23,607
2014	202	$659.85	$18,677
2015	210	$716.49	$17,173
2016	222	$759.39	$18,123

Figure 1: Consumer surplus and revenue for classic goods such as cars

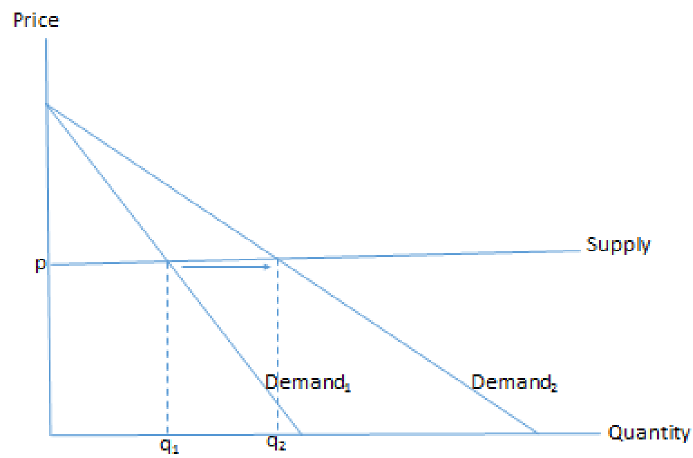


Figure 2: Consumer surplus and revenue for transition goods such as encyclopedias

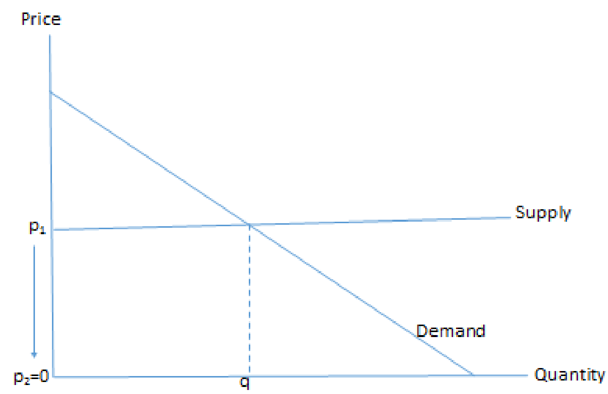


Figure 3: Consumer surplus and revenue for purely digital goods

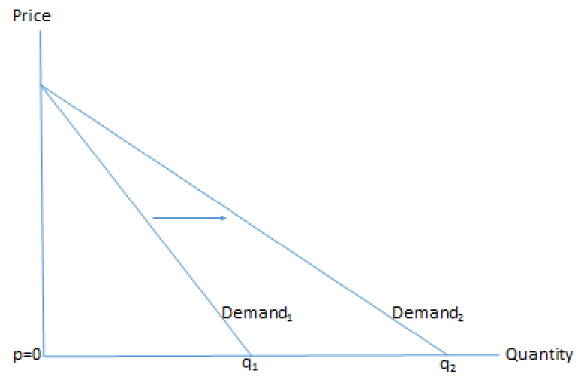


Figure 4: Example of Google Consumer Surveys

Figure 5: WTA demand curves for most widely used categories of digital goods

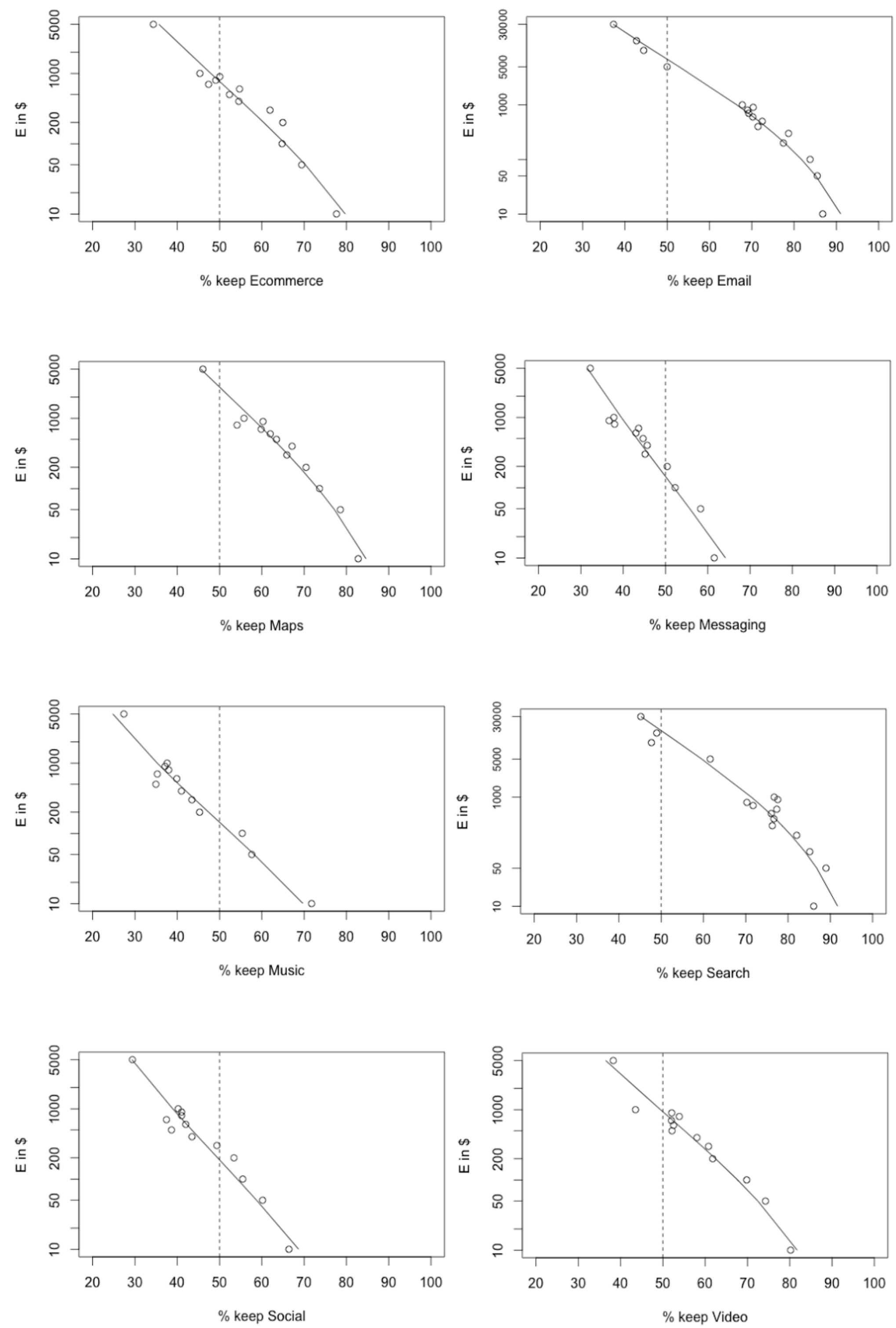


Figure 6: WTA demand curve for breakfast cereal

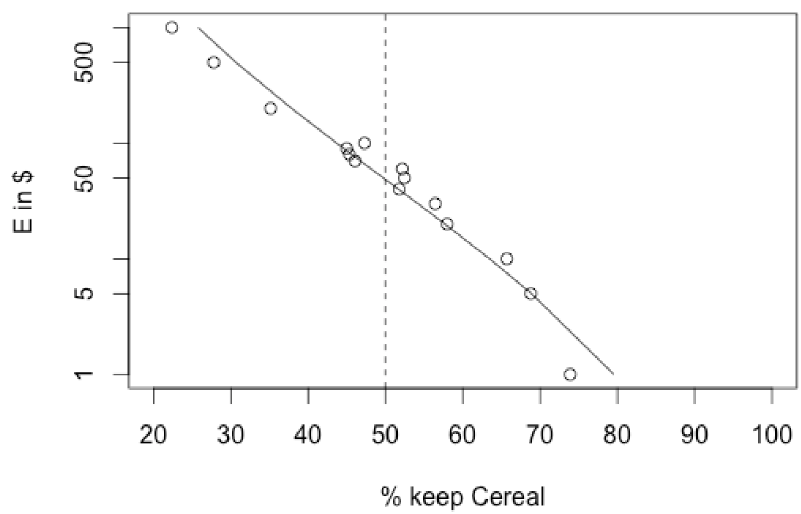


Figure 7: Exemplary best-worst scaling task

Please assume that you would have to give up access to some services or amenities for 1 year. Please consider the options below. Which of these options do you find worst and best?

	Option 1	Option 2	Option 3
	No breakfast cereal for 1 year	No access to online shopping for 1 year	Earning $500 less for 1 year
Worst option:			
Best option:			

Figure 8: Best-worst scaling estimation results

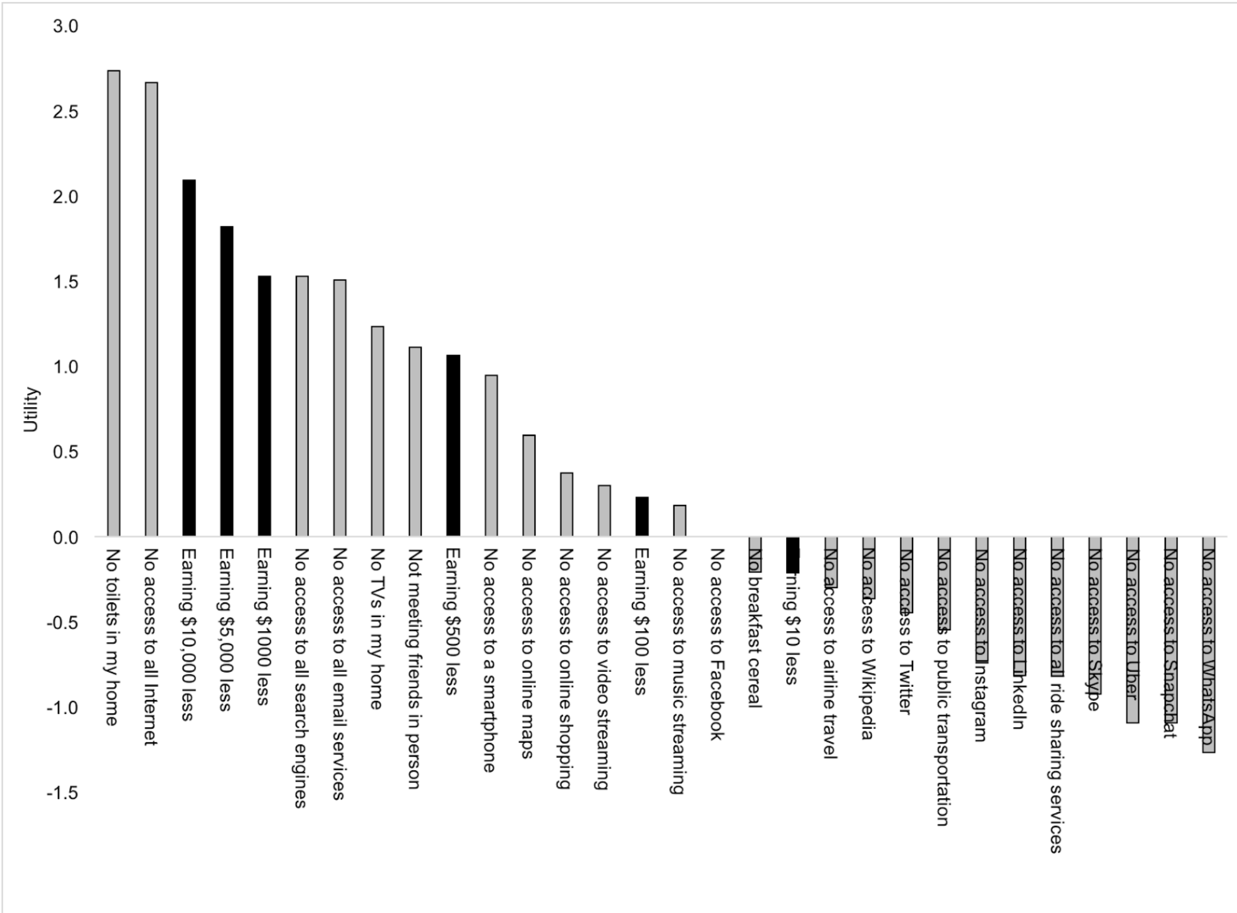


Figure 9: Assessment of hypothetical bias for Facebook

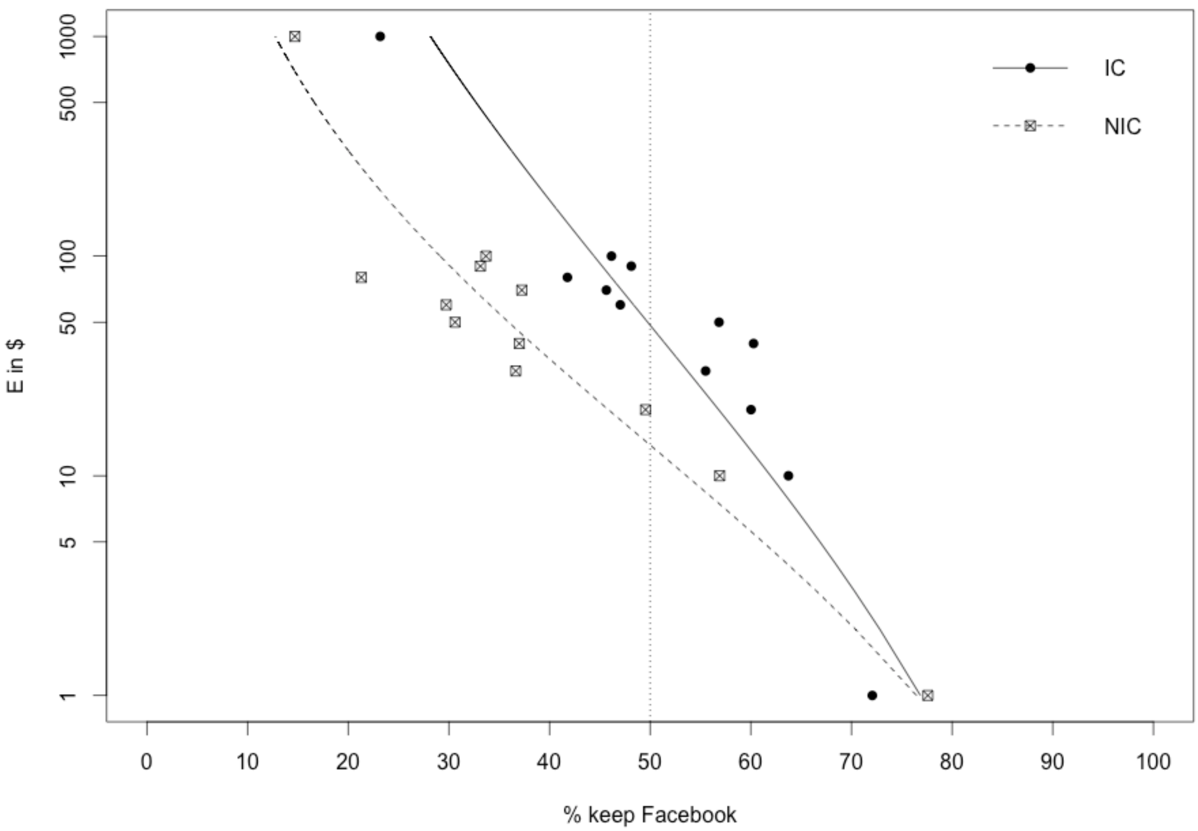


Figure 10: Comparison of BDM lottery and SBDC experiment

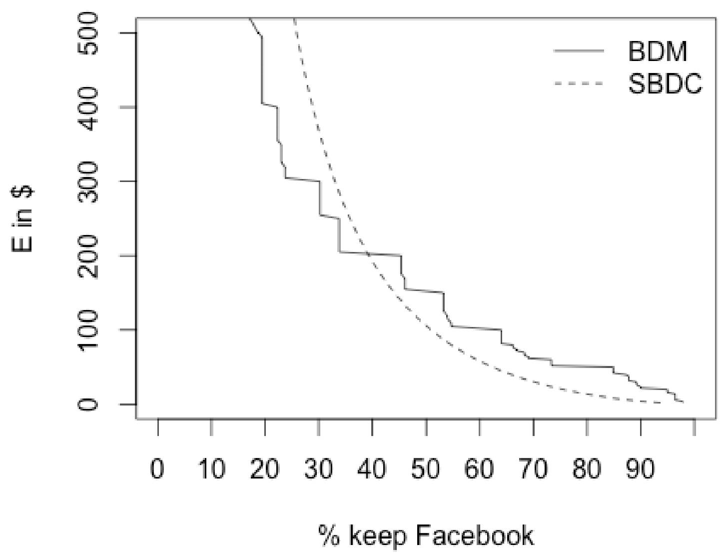


Figure 11: Effect of time on the probability to keep Facebook

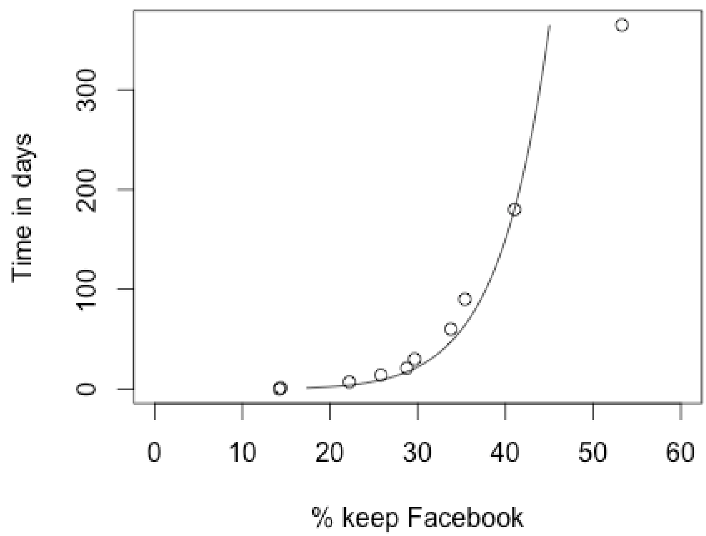


Figure 12: Effect of scale of the estimates on logit function

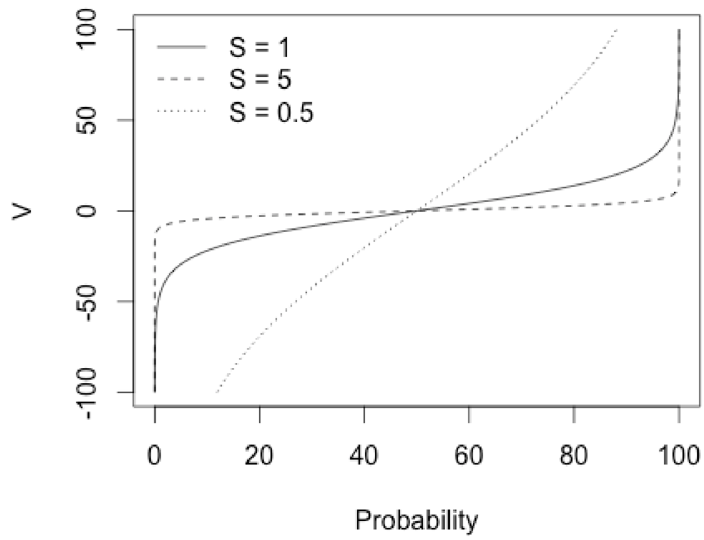
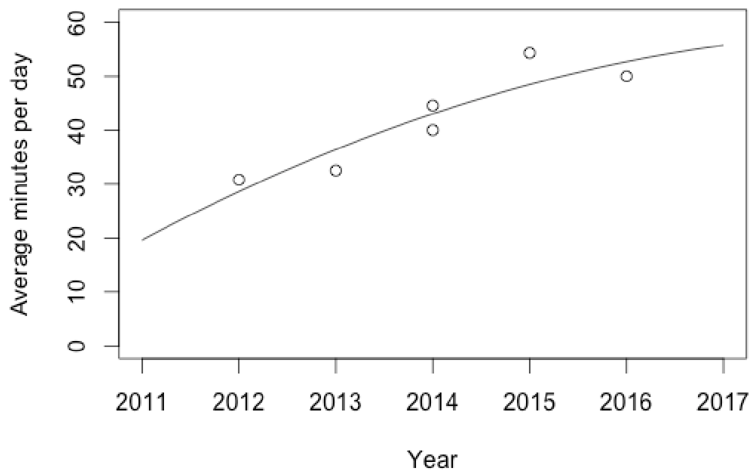


Figure 13: Facebook hours of usage over time



Appendix

Table A.1: Estimated logistic functions for most widely used categories of digital goods

	E-Commerce	Email	Maps	Messaging	Music	Search	Social	Video
log(E)	-0.352***	-0.358***	-0.321***	-0.231***	-0.324***	-0.328***	-0.285***	-0.348***
	(0.0190)	(0.0135)	(0.0198)	(0.0173)	(0.0181)	(0.0136)	(0.0178)	(0.0194)
Male	-0.110*	0.0340	-0.0565	-0.165**	-0.0703	-0.152**	-0.372***	-0.0313
	(0.0549)	(0.0545)	(0.0570)	(0.0532)	(0.0546)	(0.0557)	(0.0546)	(0.0564)
Age 18-24 (Reference)								
Age 25-34	0.154	0.221**	0.272**	-0.00158	-0.318***	0.169	-0.132	-0.0773
	(0.0884)	(0.0842)	(0.0910)	(0.0862)	(0.0882)	(0.0883)	(0.0877)	(0.0944)
Age 35-44	0.260**	0.505***	0.415***	0.258**	-0.0122	0.263**	0.172	-0.192
	(0.100)	(0.0933)	(0.103)	(0.0952)	(0.0975)	(0.0980)	(0.0982)	(0.102)
Age 45-54	0.280**	0.682***	0.368***	0.112	-0.203*	0.255*	0.206*	-0.421***
	(0.105)	(0.0996)	(0.106)	(0.102)	(0.101)	(0.102)	(0.103)	(0.106)
Age 55-64	0.524***	0.947***	0.284**	0.0397	-0.403***	0.471***	0.00909	-0.671***
	(0.106)	(0.105)	(0.107)	(0.101)	(0.105)	(0.106)	(0.104)	(0.108)
Age 65+	0.301**	1.433***	0.119	0.121	-0.499***	0.476***	0.0899	-0.578***
	(0.115)	(0.126)	(0.117)	(0.110)	(0.116)	(0.119)	(0.115)	(0.119)
Age Unknown	0.0561	0.503***	0.0766	-0.0426	-0.160	0.132	-0.249*	-0.427***
	(0.0976)	(0.0934)	(0.104)	(0.0975)	(0.0990)	(0.0964)	(0.0980)	(0.101)
Rural area (Reference)								
Urban area	-0.0154	0.216**	0.167*	0.0865	0.278***	0.175*	0.103	0.154*
	(0.0775)	(0.0753)	(0.0795)	(0.0742)	(0.0768)	(0.0769)	(0.0766)	(0.0784)
Suburban area	-0.0240	0.103	0.164*	-0.0282	0.156*	0.0791	0.0896	0.117
	(0.0740)	(0.0715)	(0.0761)	(0.0712)	(0.0742)	(0.0735)	(0.0735)	(0.0752)
Income 0-25k (Reference)								
Income 25k-50k	0.0284	0.188*	0.148	0.0521	-0.00637	0.248**	0.0623	-0.121
	(0.0943)	(0.0883)	(0.0960)	(0.0935)	(0.0936)	(0.0904)	(0.0958)	(0.0968)
Income 50k-75k	0.00192	0.280**	0.397***	-0.0406	0.0244	0.409***	-0.0124	0.0642
	(0.101)	(0.0954)	(0.104)	(0.101)	(0.1000)	(0.0983)	(0.102)	(0.104)
Income 75k-100k	0.220	0.258*	0.135	0.0926	-0.0488	0.432***	-0.0297	0.0524
	(0.131)	(0.127)	(0.136)	(0.128)	(0.132)	(0.128)	(0.133)	(0.134)
Income 100k-150k	0.0730	0.104	0.592**	-0.128	0.0449	0.418*	-0.0128	-0.140

	(0.174)	(0.178)	(0.199)	(0.176)	(0.180)	(0.179)	(0.175)	(0.180)
Income 150k+	0.523	0.120	1.201***	-0.183	-0.689*	1.098**	0.0355	-0.255
	(0.283)	(0.242)	(0.365)	(0.269)	(0.302)	(0.367)	(0.266)	(0.287)
Income Unknown	-0.0211	0.463	0.245	0.434	0.547	0.166	-0.180	-0.685*
	(0.249)	(0.254)	(0.275)	(0.278)	(0.279)	(0.268)	(0.287)	(0.277)
Income Undisclosed	0.151	-0.123	-0.0201	0.0269	-0.152	0.166	-0.162	-0.0139
	(0.177)	(0.158)	(0.171)	(0.162)	(0.175)	(0.161)	(0.174)	(0.181)
Constant	2.094***	2.273***	1.972***	1.087***	1.689***	2.631***	1.580***	2.665***
	(0.164)	(0.142)	(0.170)	(0.154)	(0.158)	(0.149)	(0.160)	(0.172)
Observations	6530	8079	6526	6600	6527	8081	6556	6572

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.2: Best-worst scaling estimation results

Good	Utility	Std. Error
No toilets in my home for 1 year	2.736	0.381
No access to all Internet for 1 year	2.665	0.378
Earning $10,000 less for 1 year	2.094	0.307
Earning $5,000 less for 1 year	1.819	0.287
Earning $1000 less for 1 year	1.531	0.267
No access to all search engines for 1 year	1.530	0.268
No access to all email services for 1 year	1.508	0.262
No TVs in my home for 1 year	1.234	0.253
Not meeting friends in person for 1 year	1.111	0.245
Earning $500 less for 1 year	1.064	0.245
No access to a smartphone for 1 year	0.949	0.240
No access to online maps for 1 year	0.597	0.232
No access to online shopping for 1 year	0.377	0.221
No access to video streaming for 1 year	0.301	0.221
Earning $100 less for 1 year	0.231	0.215
No access to music streaming for 1 year	0.183	0.221
No access to Facebook for 1 year	(0.000)	
No breakfast cereal for 1 year	-0.208	0.210
Earning $10 less for 1 year	-0.211	0.212
No access to airline travel for 1 year	-0.299	0.212
No access to Wikipedia for 1 year	-0.362	0.212
No access to Twitter for 1 year	-0.446	0.210
No access to public transportation for 1 year	-0.545	0.211
No access to Instagram for 1 year	-0.729	0.214
No access to LinkedIn for 1 year	-0.816	0.212

No access to all ride-sharing services for 1 year	-0.818	0.213
No access to Skype for 1 year	-0.921	0.211
No access to Uber for 1 year	-1.092	0.212
No access to Snapchat for 1 year	-1.093	0.213
No access to WhatsApp for 1 year	-1.265	0.213

Figure A.1: Example of Incentive Compatible (IC) Questionnaire for Facebook SBDC question (for E = $80)

Would you prefer to keep access to Facebook or go without access to Facebook for 1 month and get paid $80?

We want to reward you for thinking carefully about this question. Therefore, we will randomly pick 1 out of every 200 respondents and her/his selection will be fulfilled:

- If you choose "Keep access to Facebook" you can keep using Facebook as before. However, you will not receive the $80 in cash.
- If you choose "Give up Facebook and get paid $80" you will receive the $80 in cash, provided that you do not access Facebook for 1 month. Facebook collects the date and time when you have last used your account. Given your permission, we can access this time with an app (e.g., see this link for an example). In order to get your permission, we will contact you via email. You can revoke this permission at any time.

Therefore, it is in your best interest to think carefully about how valuable you find Facebook.

What is your decision: Would you prefer to keep access to Facebook or go without access to Facebook for 1 month and get paid $80?

○ Keep access to Facebook
○ Give up Facebook and get paid $80

Proceed to next page

Robustness of results: Selection bias

Table A.2: Estimation results of binary logit model comparing Peanut Labs (non-incentive compatible group) and GCS

	beta	Std. Error	z	Sig.
(Intercept)	0.579	0.114	5.091	<0.001
log(E)	-0.374	0.029	-12.686	<0.001
Google	0.002	0.168	0.011	0.991
Google*log(E)	0.031	0.043	0.715	0.474

Figure A.2: Assessment of selection bias