NBER WORKING PAPER SERIES

INFORMATION AND THE WELFARE BENEFITS FROM DIFFERENTIATED PRODUCTS

Imke Reimers
Christoph Riedl
Joel Waldfogel

Working Paper 33401
http://www.nber.org/papers/w33401

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2025

The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

Information and the Welfare Benefits from Differentiated Products
Imke Reimers, Christoph Riedl, and Joel Waldfogel
NBER Working Paper No. 33401
January 2025
JEL No. L15, L82

ABSTRACT

Differentiated product consumption choices made without full information can lead to welfare losses from regret and missed opportunities, but a lack of post-purchase usage data has prevented their exploration. Using novel data on individual ownership and post-purchase usage of video games, we explore both the potential welfare benefits of full information prior to purchase and the ability of contemporary prediction technology to produce these gains. We find large potential gains: Among currently owned games, fully informed consumers could achieve 90 percent of their status quo playtime with 40 percent of current expenditure; and current expenditure reallocated among all available games could double status quo playtime. We develop a tractable model of consumer choice among bundles based on hours of playtime relative to overall spending, which we implement using both a Cobb Douglas calibration and a logit model of bundle choice. Full information would raise consumer surplus by more than the value of status quo expenditure; and it would reduce expenditure by half. Consumers heeding sophisticated, personalized predictions would obtain roughly 40 percent of these welfare benefits with a fifth less spending.

Imke Reimers
Cornell University
Dyson School of Applied Economics
and Management
SC Johnson College of Business
351C Warren Hall
Ithaca, New 14853
United States
imke.reimers@cornell.edu

Christoph Riedl
Northeastern University
c.riedl@neu.edu

Joel Waldfogel
Frederick R. Kappel Chair in Applied Economics
3-177 Carlson School of Management
University of Minnesota
321 19th Avenue South
Minneapolis, MN 55455
and NBER
jwaldfog@umn.edu

1 Introduction

The availability of a wide variety of differentiated products delivers substantial benefits to diverse consumers. Yet, the varied features of products – and varied tastes of consumers – can make it hard for consumers to know which products they might like. Some current purchases may disappoint their buyers, and some worthwhile options may be missed. If consumers had full information about all products prior to purchase, then differentiated product purchases might deliver substantially more utility. These problems of regret and missed opportunities can arise in a wide swath of the economy, including markets for information goods and apparel, and even in labor, healthcare, education, and housing markets.[1]

New prediction technologies and better data have enabled personalized recommendations that may improve consumers' pre-purchase information, allowing some of the gains. These considerations raise two questions. First, how much welfare is forgone in status quo differentiated product consumption choices, relative to full information? Second, how large an improvement over status quo welfare could sophisticated, personalized predictions deliver, if consumers heeded them?

Assessing the extent of regret and missed opportunities in current consumption decisions has traditionally been difficult, for three reasons. First, we do not normally observe post-purchase usage. Hence, there is little empirical basis for evaluating the ex post welfare delivered by current consumption choices. Second, data on post-purchase usage of the products people have purchased is not sufficient. A complete answer also requires information on how much people would have used the products they did not purchase. Third, beyond these

[1] Relevant contexts include markets for "experience" goods, as in Nelson (1970). Skelton and Allwood (2017) provide systematic evidence that clothing, takeout food, sports & exercise equipment, and kitchen gadgets are among the most regretted purchase categories. Einav et al. (2023) show evidence consistent with regret in markets for subscription services. There is evidence of regret in other contexts as well. See https://stradaeducation.org/value/do-you-regret-your-college-choices/ on education, https://realestate.usnews.com/real-estate/articles/have-buyers-remorse-with-your-home-heres-what-to-do on housing, and https://www.forbes.com/sites/bryanrobinson/2022/12/01/5-reasons-for-boomerang-employees-and-the-great-regret-in-employment/ on labor markets.

challenging data requirements, we also need a tractable way to model consumer choice and welfare in an environment with many possible bundles and imperfect pre-purchase information.

We surmount the data challenge with unusual individual-level data on product usage. Not only do we observe which consumers purchased each of the products, we also see the amount of time they spent using them. Our data, from the video game platform Steam, include information on 50,000 consumers' cumulative usage of 100 games. Using only data on owned games, we document that consumers commonly make seemingly regrettable choices, i.e., they purchase products they use very little and which they recommend that others not purchase.

Still more gains would be possible if consumers were aware of products they did not purchase but would have enjoyed. Quantifying the potential benefits from games that users do not currently own – addressing the second challenge – requires additional steps. First, we need a plausible characterization of the *truth*, i.e., the realized playtime that each potential purchase would deliver. Second, to estimate the efficacy of contemporary prediction technology, we also need personalized predictions resembling those used in practice. We observe the true realized playtimes directly for the owned games, and we use these to create predictions of playtime for each user and game using matrix factorization. Then, armed with both predictions and information on the distribution of the associated prediction errors for owned games, we simulate the "truth" for unowned games as their predicted playtime plus an error that reproduces the correlation between the predicted and realized playtime for owned games.[2]

We show in descriptive analyses that consumers could experience large savings by avoiding disappointing status quo purchases. Among currently owned games, fully informed consumers could achieve the vast majority (90 percent) of their status quo playtime with less than half (40 percent) of status quo expenditure. The full effect of better information

[2] Our results are robust to various alternative ways of modeling true playtime for unowned games.

includes the benefits of both avoiding regret and discovering previously-missed opportunities. Allowing for both mechanisms, we find that, on average, fully informed consumers could achieve a 94 percent increase in realized playtime with current expenditure; they could conversely purchase current playtime with one fifth of current expenditure. Contemporary predictions, if heeded, would allow a 36 percent increase over the status quo playtime with current expenditure, or the achievement of current playtime with 38 percent less expenditure. These estimates, which hold current expenditure constant, are robust to various assumptions about true playtimes, prices, and returns. They suggest large welfare gains but do not enable direct calculation of their value.

Addressing the third challenge – how to model consumer preferences and welfare – requires an implementable economic framework relating pre-purchase information to consumer choice over bundles of games. We simplify the multidimensional problem by modeling consumers' utility as depending on cumulative hours of game playtime, less the expenditure on the games.[3] Our approach makes the strong assumption that hours of different games deliver the same marginal utility, and we provide empirical justification for this approach, as well as robustness checks, below.

The quality of pre-purchase information available to consumers affects the consumers' ex ante choice problem through a budget constraint running from all other goods on the y-axis to hours of playtime on the x-axis. If consumers were ignorant – they knew the average price of playtime but nothing about the playtime delivered by each game – their expected budget constraints would be linear (based on the average hourly price of playtime). A fully informed consumer, by contrast, has a "bowed-out" budget constraint reflecting the purchase of the lowest price-per-hour games first. Intermediate levels of information – such as the status quo, or the availability of sophisticated predictions – deliver intermediate

[3] Our simplification recalls Chu et al. (2011)'s recasting of mixed bundling as a problem involving the number of products purchased.

budget constraints. Given their utility functions and the budget set associated with their pre-purchase information, consumers choose a level of expenditure to maximize attainable utility; and true playtime ensues.

This framework allows us to address our two main questions, measuring the welfare benefits from full information and the extent of the possible effect delivered by sophisticated predictions. We use the framework both to calibrate a Cobb Douglas model of demand for time spent playing, and to estimate conditional logit models in which the consumer prefers their chosen bundle of games (and hours) and expenditure to non-chosen alternative bundles. Our results from these models echo the descriptive results. Relative to the status quo, full information raises average consumer surplus (CS) by about 130 percent of status quo expenditure while reducing expenditure by a half. Nearly 40 percent of full information's effect comes from avoiding regrettable purchases; the remainder arises from taking advantage of otherwise-missed opportunities. Heeding personalized, sophisticated predictions also allows around 40 percent of the overall benefit of full information while reducing expenditure by two fifths. The sophisticated predictions are better than status quo choices for 88 percent of the users in our data, while 12 percent are better off in the status quo.

The paper proceeds in five sections. Section 2 discusses the relevant academic literatures and the product-market context; we also provide contextual evidence of ex-post regret. Section 3 introduces our tractable model of the effect of pre-purchase information on a consumer's choice among product bundles. Section 4 describes the data used in this study as well as both predictions of playtime and our main and alternative measures of true realized playtime. Section 5 presents descriptive evidence suggestive of large welfare gains from better pre-purchase information. We begin with exercises that do not rely on predictions, and we show that observed purchases include many apparently-disappointing choices. We then use our predictions and measures of true playtime to document the additional playtime – and, conversely, reduced expenditure – that full information and sophisticated predictions could enable. Section 5 also shows that our descriptive findings are robust to alternative measurement assumptions and presents evidence supporting the use of aggregate hours in the utility function. Section 6 presents our empirical structural model of the budget constraint and our model of utility, in which consumers choose bundles of games to maximize their utility from playtime, and Section 7 shows welfare effects of full information and sophisticated predictions, as well as the components of these effects arising from reduced regret. In addition, Section 7 discusses the heterogeneous effects of information across users and games, shows that similar results obtain when we allow demand parameters to differ by user and game type, and provides measures of the welfare benefit from a sequence of increasingly sophisticated predictions. We conclude in Section 8.

2 Background

2.1 Relevant literatures

Our broad question – about the effect of pre-purchase information on the welfare consequences of differentiated products markets – is related to five literatures. First, our work is related to studies of inefficiencies arising in differentiated products markets (Spence, 1976; Dixit and Stiglitz, 1977; Mankiw and Whinston, 1986; Anderson et al., 1995). The literature has traditionally focused on supply-side challenges associated with the number of entering products; we are instead concerned with the demand-side efficiency consequences of information, given the available products.

Second, our study is relevant to work on the welfare benefit from large numbers of new products, including those associated with "the long tail." See Anderson (2007), Brynjolfsson et al. (2003), Quan and Williams (2018) and Aguiar and Waldfogel (2018) on the welfare benefits of product variety and Waldfogel (2007) for evidence on differentiated product markets and diverse consumers. Like Chu et al. (2011) and Crawford and Yurukoglu (2012), we estimate welfare consequences of bundle choices.

Third, the part of our paper documenting the efficacy of prediction is related to research on the effect of pre-purchase information on purchase decisions. One strand of existing research documents effects of non-personalized information, which have been shown to affect purchase decisions and to improve welfare (Reinstein and Snyder, 2005; Chevalier and Mayzlin, 2006; Reimers and Waldfogel, 2021). Some papers document apparent ex post mistakes (Allcott, 2013; Miravete, 2003). More recent papers show the effects of specific examples of personalized recommendations on purchase behavior and welfare (Sun et al., 2024; Donnelly et al., 2023; Kaye, 2023; Wu et al., 2023). It is clear from prior work that pre-purchase information can affect purchase decisions and welfare; what we contribute is quantification of both inefficiencies of current consumption and the share of the potential benefits from full information put in reach by contemporary predictions.

Fourth, we draw on the literature on recommender systems (Koren et al., 2009; Bobadilla et al., 2013; Lee and Hosanagar, 2021; Koren et al., 2021). Specifically, we use collaborative filtering which leverages past consumer behavior such as purchases or product ratings for observed consumer-product pairs to estimate usage of consumer-product pairs that are not observed. The matrix factorization approach we use is representative of prediction approaches used in practice (Koren et al., 2021); and it has outperformed more complex functions such as neural networks (Rendle et al., 2020). For example, matrix factorization predictions were used to win the Netflix Prize (Koren et al., 2009).

Finally, there are substantial literatures on various aspects of video games, including the complementarity between consoles and games (Lee, 2013), the potential impacts of video games on social outcomes (Ward, 2010), and the relationship between work hours and video games (Aguiar et al., 2021).

2.2 Industry context

Video games attract substantial amounts of entertainment spending, as well as time use. The video game industry generated $347 billion in worldwide revenue in 2022, making it substantially larger than the movie and music industries combined.[4] The US Bureau of Labor Statistics reports that Americans spent an average of 34.2 minutes per day playing video games during 2022.[5] Between 2014 and 2017, US men between the ages of 21 and 30 spent an average of 3.9 hours per week playing video games (Aguiar et al., 2021). US consumers spent $47.5 billion on video game content during 2022 while consumer spending overall was $9.8 trillion.[6] Hence, video games accounted for 0.49 percent of household spending, which we use to inform our expenditure share estimate in the Cobb Douglas analyses.

Video games are played on game consoles (such as the Nintendo Switch or Sony PlayStation), on phones, or on computers, where games are downloaded from digital video game distribution platforms. One of the largest such platforms, Steam, provides the setting for our analysis. Operated by Valve Corporation and founded in 2003, Steam had 33 million concurrent peak users during 2023.[7] Steam offers over 73,000 games, and revenue from game sales on Steam was $8.8 billion in 2022, about 20 percent of total US spending on video games.[8]

[4] See https://www.statista.com/topics/868/video-games/topicOverview. Global recorded music revenue was $31.2 billion in 2022. See https://www.statista.com/statistics/272305/global-revenue-of-the-music-industry/, while global movie revenue was estimated at $93.4 billion. See https://www.ibisworld.com/global/market-size/global-movie-production-distribution/. Global movie box office alone was $26 billion. See https://www.imdb.com/news/ni63899899/.

[5] See https://www.statista.com/statistics/502149/average-daily-time-playing-games-and-using-computer-us-by-age/.

[6] See https://www.statista.com/statistics/252457/consumer-spending-on-video-games-in-the-us/ and https://www.bls.gov/opub/reports/consumer-expenditures/2022/home.htm.

[7] See https://www.statista.com/topics/4282/steam/topicOverview.

[8] See https://www.statista.com/statistics/547025/steam-game-sales-revenue/.

2.3 Evidence of ex-post regret

The large number of games available on the Steam platform makes it difficult for consumers to know which products they might find appealing. It is not surprising, therefore, that expressions of regret are common from video game users.

Numerous social media sites feature discussions of games that consumers regret buying. A Reddit thread entitled "What's one game you regret buying?" elicited 1,700 comments. The top (most upvoted) reply was, "Probably 70% of my steam library."[9] Similar comments are shared at Quora; and a substantial genre of YouTube videos describes games that users regret buying.[10] Together, those comments suggest that purchase errors are not only possible but common in this context.

User recommendations provide additional, and more systematic, evidence of varying post-purchase reactions to games. Steam users leaving feedback can "recommend" or "not recommend" a game. Among the 50,000 users in our sample, 40,370 leave reviews; and 18.6 percent of the reviews do "not recommend" games to other users. The prevalence of regretted purchases suggests that consumers lack full information prior to purchase and, by extension, that better pre-purchase information could raise welfare.[11]

3 Theory

We are interested in analyzing how better pre-purchase information would affect consumers' choices among high-dimensional bundles of products. In our context, consumers are choosing which bundles, from among 100 games, to own. These decisions depend on their utility functions and their information about games. We discuss each in turn below.

[9] See `https://www.reddit.com/r/gaming/comments/12frdsr/whats_one_game_you_regret_buying/`

[10] See, for example, "Video Games I Regret Buying" (`https://www.youtube.com/watch?v=l4f8CmcLJPk`) or identically titled video (`https://www.youtube.com/watch?v=WoB80aEQsnE&t=16s`).

[11] We collected these data from `https://steamcommunity.com/profiles/[steamid]/recommended` during June of 2024.

3.1 Consumer utility

In general, consumers would have some utility over the bundle of games, less the utility of money paid for the games:

$$u_{ij} = U(\mathbb{1}_{i1}, \ldots, \mathbb{1}_{iJ}) - \sum_{j \in J} p_j \mathbb{1}_{ij},$$

where $\mathbb{1}_{ij}$ is an indicator that is 1 if individual i owns game j, J is the full set of available games, and $P_i = \sum_{j \in J} p_j \mathbb{1}_{ij}$ is the spending required to purchase the bundle of owned games. Ideally, the utility function $U()$ would allow for both varying marginal utilities across games, as well as substitutability among games. With 100 games in the choice set, there are 2^{100} possible bundles, so to make progress, we need some simplification. The simplification we employ is to assume that users derive utility from games according to the hours of playtime that the games deliver, or

$$u_{ij} = U\left(\sum_{j \in J} h_{ij} \mathbb{1}_{ij}\right) - \sum_{j \in J} p_j \mathbb{1}_{ij}, \tag{1}$$

where h_{ij} denotes the amount of time that consumer i would use product j. This approach entails different marginal utilities of ownership across games in the sense that they are proportional to how many hours of playtime the respective games deliver. Depending on the functional form of U, the approach can also allow for game substitutability via diminishing marginal utility of the amount of playtime that user i's chosen bundle would deliver, $H_i = \sum_{j \in J} h_{ij} \mathbb{1}_{ij}$. Perhaps the strongest implicit assumption embodied in this approach is that utility of games depends on hours in way that is identical across games. We provide empirical support for reducing the bundle choice to an hours choice in Section 5.4 below.

3.2 Pre-purchase information and the opportunity set

Given what a consumer knows about products prior to purchase, the consumer faces a budget constraint describing how expenditure delivers playtime. To derive the budget constraints, it is helpful to begin with two extreme cases, one in which consumers have no information about individual games and another in which they have full information. In the "no information" case, consumers know the average amount of playtime they obtain per dollar spent, which we term ρ_i, but not the particular realized value for each game. Then a consumer spending a total of P_i would expect to receive $H_i = \rho_i P_i$ hours of playtime. We illustrate a budget constraint resulting from random rankings (which we term r^ϵ) in the dashed line in Figure 1, which plots the cumulative amount of playtime (x-axis) against money available for all other goods (y-axis). The consumer would face a linear expected budget constraint, with the same expected amount of money per hour $^1/_{\rho_i}$ for each purchased game. While each budget constraint realization depends on the random rank order draw (r^ϵ), the budget constraints will be linear in expectation across draws.

At the other extreme, consumers have full information on the hours they would play each game (h_{ij}^T). When armed with full information, consumer i knows that game j would deliver h_{ij}^T of playtime at a price of p_j. Ordering products by ascending values of $^{p_j}/_{h_{ij}^T}$, a ranking we term r^T, delivers the maximally expansive budget constraint for the consumer. The outer budget constraint in Figure 1, drawn curved to reflect a continuous approximation, represents this full-information case.

The shape of the consumer's full information budget constraint depends on the variability of their price per hour of playtime across games, $^{p_j}/_{h_{ij}^T}$. If all games delivered the same hours of playtime per dollar spent, then the full information budget constraint would be linear and indistinguishable from the no-information budget constraints. The greater the variance in a user's price per hour, the more bowed out is their full information budget constraint.

Any ranking of products based on something other than the true $^{p_j}/_{h_{ij}^T}$ shrinks the opportunity set relative to the full information case. For example, if consumers had access to sophisticated pre-purchase predictions, they would not know h_{ij}^T in advance but rather a prediction h_{ij}^P containing error. A consumer heeding the personalized prediction would rank-order games by price per predicted hour of use, or ${}^{p_j}\!/_{h_{ij}^P}$, producing a sophisticated-prediction budget constraint. Provided that the prediction is better than random, this delivers a realized budget constraint that lies somewhere between the two extremes of ignorance and full information; and the inner, curved budget constraint in Figure 1 illustrates such a scenario. The deviation of the prediction-based budget constraint from the full information budget constraint is larger, the less accurate the prediction.

We need a way to characterize consumers' status quo budget constraints. Consumers in the status quo could have a range of possible budget constraints that would generally fall short of full information. A ranking according to r^T maximizes the playtime that each expenditure delivers, whereas a random ranking (r^ϵ) produces a linear budget constraint in expectation; and the linear budget constraint is not the most extreme possibility. Instead, a consumer's information could be "worse than random" in the sense that their ranking could be negatively correlated with r^T. At an extreme, a consumer who ranked products according to $-r^T$ would attain minimum playtime with any level of expenditure. To accommodate the possibility that information might vary from full information to worse than random, we create an index that weights the random and full information rankings via

$$I(\kappa_i) = \kappa_i r^T + (1 - |\kappa_i|) r^\epsilon, \tag{2}$$

where $\kappa \in [-1, 1]$, and the consumer's resulting ranking r^I is in order of $I(\kappa_i)$.

The shape of a consumer's budget constraint depends on the size and sign of κ_i. If the consumer possessed full information in the status quo, then $\kappa_i = 1$; and the consumer's status quo budget constraint would lie on the full information budget constraint. If the status quo consumer has better-than-random but less-than-full information, then $0 < \kappa_i < 1$ and the average status quo budget constraint is less bowed out. If $\kappa_i = 0$, the consumer's budget constraint is linear with a slope based on the average price per hour across games, as depicted by thedashed linear budget constraint in Figure 1. Finally, if $-1 \leq \kappa_i < 0$, then the consumer's knowledge gives rise to a budget constraint that is "bowed in" toward the origin.

Depending on the consumer's information (κ_i) relative to the accuracy of the personalized predictions, a consumer's status quo budget constraint could lie inside, or outside, of the budget constraint based on these predictions. Indeed, it is an empirical question whether – and for what share of users – contemporary predictions contain better information than consumers possess in the status quo.

Given the budget constraint associated with their pre-purchase information, the consumer chooses a point such that their marginal rate of substitution (MRS) equals the slope of the expected budget constraint. Figure 1 illustrates the utility-maximizing hours choices with a full information and a less informed budget constraint. In what follows, we develop empirical characterizations of the status quo, full information, and prediction-informed budget constraints, along with models of utility that deliver both status quo and counterfactual choices and their welfare effects.

4 Data and playtime measures

4.1 Data

The main data for this study include information on 50,000 Steam users and 100 popular Steam games. For each user, we observe which of these games they own, as well as the cumulative number of hours they have spent playing each of the games, as of data collection between May 9 and 16 of 2021.[12] The underlying data include 192,137 users, who collectively owned 33,844 distinct games. From these, we chose the 100 most popular games (by number of users who bought the game) with positive prices. To ensure sufficient usage history data, we restrict attention to users who purchased at least 20 of those popular games; and we randomly selected 50,000. We collected price data for each game from `https://steampricehistory.com` on November 15, 2023. We obtain a single price for each game by averaging prices over time between January 2015 and May 2021.

Our users spend an average of $508.10 in total and own an average of 33.64 games. These games provide them with 2,166.4 hours of cumulative playtime; and users play each game an average of 64.4 hours. These averages take into account two features of the environment. First, our main analysis assumes that all users paid the average price for each game, even though games on Steam are sometimes discounted. In Section 5.3 we explore the sensitivity of our results to the possibility that users obtained varying shares of their games at no charge.

Second, Steam allows users a two-week window to return games played less than 120 cumulative minutes. The playtime data thus include times spent briefly playing games that the users ultimately neither purchase nor continue playing. Leaving these in the sample would lead us to overstate welfare effects of better information, as the status quo holdings would appear to include bad games that the users did not actually purchase. We deal with this by eliminating potentially returned games from the sample in a way that is informed by aggregate return tendencies: Industry sources indicate that between 5 and 8 percent of purchased games are returned under Steam's policy.[13] In our data, 6.5 (the midpoint between 5 and 8) percent of game purchases with nonzero playtimes are played less than 23 minutes. We mimic the true return process by excluding the game purchase instances in which the games were played less than 23 minutes. We explore the consequences of other

[12] See `https://developer.valvesoftware.com/wiki/Steam_Web_API`, which we used to obtain lists of owned games and their playtime for players with publicly visible profiles.

[13] See `https://newsletter.gamediscover.co/p/game-refunds-and-the-hidden-costs`.

time cutoffs (between 0 and 120 minutes) in Section 5.3 below.

We have two kinds of additional data, which we use for some prediction models and for extensions to our logit demand estimations. First, we have game characteristics. In addition to price, we observe each game's genre. Games in the "action" genre are by far the most common (71 percent). Second, we know when 86.8 percent of users joined the Steam platform. Join years vary from 2003 to 2020, with a median of 2010 and an inter-quartile range from 2007 to 2013. We show in Section 7.2 that our results remain when we allow different game types to enter the utility function separately, as well as when we restrict the sample to users who joined during particular years and who therefore have had similar amounts of time to accumulate playtime.

4.2 Modeling true playtime and realistic predictions

Answering our research questions requires two kinds of playtime measures. First, we need sophisticated predictions of playtime h_{ij}^P for each player i and game j that are consistent with the kinds of predictions employed in practice. Second, we need measures of true, realized playtime h_{ij}^T for each user and game, including those not currently owned. This section discusses how we use matrix factorization to obtain playtime predictions, and how we create measures of true playtime by adding realization errors to the predictions for unowned games.

4.2.1 Predictions of playtime from matrix factorization (h_{ij}^P)

Our users own a third of the 100 games on average, so the data matrix for generating predictions is sparse.[14] One can imagine a variety of approaches to filling in the prediction matrix. Our preferred prediction approach is to create a collaborative filter using matrix factorization, and we show below that our approach outperforms other prediction techniques.

[14] This has been formulated as a matrix completion problem, in which missing elements of the user-item matrix have to be predicted from limited historical data as not every user has interacted with every item (Jannach et al., 2016).

Our implementation of the collaborative filter follows the approach of Koren et al. (2009). We employ a matrix factorization model that maps both users and products to a joint latent factor space with $k = 100$ dimensions. Each product j is associated with a vector $\mathbf{m}_j \in \mathbb{R}^k$, and each user i is associated with a vector $\mathbf{n}_i \in \mathbb{R}^k$. The elements in $\mathbf{m}_j$ ($\mathbf{n}_i$) capture the extent to which a product (user) possesses those latent factors. We estimate the fitted values of the log playtime that a user would derive from a given game from the inner products in that latent space, $\hat{h}_{ij} = \mathbf{m}_j^T \mathbf{n}_i$.

Estimation of the predicted playtime values proceeds in two steps. First, we tune the hyperparameters of our prediction model to minimize the regularized squared error between observed hours and the product $\mathbf{m}_j^T \mathbf{n}_i$ on the set of owned games, without overfitting. These hyperparameters are the number of iterative model fitting steps, the step size by which parameters change, and the penalty for parameters that are very large or very small.[15] We tune the hyperparameters on a training set that includes all but one owned game per user, which we save for the test set.[16] From the test set, we retain 10 percent as a final validation data set that does not inform the chosen hyperparameters.

In the second step, we refit the model using the hyperparameters with the lowest validation-set error from above. We do this ten times, holding out a different tenth of the sample each time. This gives us out-of-sample playtime predictions for every owned and unowned user-game combination.

Our latent factor estimation uses stochastic gradient descent optimization (Funk, 2006). Although the modeling approach includes no observable characteristics of games nor users, the large number of latent factors implicitly captures these types of variation. In the video game context, the 100 latent factors we estimate might capture obvious dimensions such as

[15] We fine-tune the model's hyperparameters by independent uniform random sampling of parameter values (Bergstra and Bengio, 2012).

[16] As is typical in the recommender literature, we perform leave-one-out cross validation. Because there is no natural order to individuals' interactions with products, we selected one game at random for each individual as test data.

adventure games, role playing games, and first-person shooter games. They may also capture less well-defined dimensions such as pace of game play (real time vs. round based), the visual style of the game (realistic vs. cartoon), or dimensions that cannot be interpreted at all.

How well do our out-of-sample predictions perform? Table 2 shows the root mean squared errors (RMSE) from a sequence of prediction approaches, from the global average (using the average value of playtime across users and games as the common prediction) to matrix factorization with 100 factors.[17] The RMSE measures for the out-of-sample validation sets run from 0.722 with the global average to 0.607 with our preferred, "sophisticated" prediction approach.[18] Adding simple user characteristics, as well as country indicators, improves only slightly on the global average, to 0.720. Using simple game attributes but no user characteristics improves RMSE to 0.701; adding game fixed effects improves RMSE to 0.648, while using all user and game observables in a random forest brings the RMSE to 0.640. Collaborative filtering approaches using matrix factorization improve substantially on most of these. Using only 5 factors delivers RMSE of 0.647, 10 factors give an RMSE of 0.639, and 50 and 100 factors deliver 0.611 and 0.607, respectively.

Our preferred approach substantially outperforms the alternative techniques. Section 7.4 compares the welfare gains delivered by progressively more accurate prediction approaches.

4.2.2 Measures of true playtime

In addition to playtime predictions h_{ij}^P, we also need measures of true playtime h_{ij}^T. We directly observe realized playtime for the owned games; the additional task is to create realistic measures of truth for the game-user cells where playtime is not observed. One might

[17] We discuss all prediction approaches in more detail in Appendix Section A.

[18] Achievable RMSE values are often quite compressed (Koren, 2009) but there is evidence that even small improvements in RMSE terms can have a significant impact on the quality recommendations. To put this number in perspective: The equivalent improvement of personalized recommendations via matrix factorization ($k = 50$) over product-level averages on the Movielens 100K dataset – a common benchmark dataset used in the recommender literature – and predicting star ratings on a 1-5 scale is 8.1 percent (Adomavicius and Zhang, 2012).

be tempted to use the sophisticated predictions as estimates of true playtime, but doing so would make predictions artificially appear to deliver the benefits of full information. Instead, we need to realistically model the errors in the predictions.

Our main approach estimates h_{ij}^T for the unowned games as the prediction plus an error. We choose the errors from the empirical distribution for owned games as follows. First, we sort all observations by the prediction h_{ij}^P. Then, for unowned games, we use the realized error associated with the owned observation with the next-highest value of h_{ij}^P. This approach retains the true realized values for owned games. Because of the way that the errors are assigned to predicted playtimes h_{ij}^P, we reproduce the same correlation of predicted playtime and the error for both owned an unowned games.[19] In addition to the main approach, we also explore four alternative approaches using empirical and parametric error distributions in Section 5.3.

We report summary statistics for prices as well as realized and predicted playtimes in Table 1, which reports averages of user-by-game observations for owned games, unowned games, and total games. Prices average $15.22 per game, with little difference between owned and unowned games. By contrast, owned games deliver both more predicted and realized playtime than unowned games. Using the true measure, owned games deliver an average of 6.65 log minutes, while unowned games deliver an average of 5.80, or 64.4 and 39.9 hours, respectively. Predicted playtimes are similarly different between owned and unowned games. This suggests that status quo consumers on average have some information.

5 Descriptive welfare evidence

Using our measures of true and predicted playtime, we document the additional hours of playtime that status quo expenditures could buy if users had better information. We do this

[19] Unlike linear regressions, which produce errors that are orthogonal to predictions, the deviations of realized playtime from the matrix factorization predictions can be correlated.

in four parts. First, Section 5.1 analyzes owned games, for which we observe true realized playtime directly; and we calculate measures suggestive of regrettable choices. Second, Section 5.2 uses both owned and unowned games to calculate the additional playtime – or reduced expenditure – that better pre-purchase information would allow. Third, Section 5.3 shows that the descriptive results are robust to different measurement assumptions about game usage and prices. Finally, Section 5.4 shows that our results are robust to allowing users to value playtime differently across games.

5.1 Analysis of owned games

The data on games that users currently own provide a useful first glimpse into the regrettable nature of status quo consumption. The analysis of owned games has the additional feature that it can proceed without reliance on our estimates of truth for unowned games.

Consumers in our sample spend an average of $508.1 to purchase games delivering 2,166.4 hours of playtime. Here, we calculate the maximum hours potentially available to a user at any level of expenditure by ordering the games they had purchased by p_j/h_{ij}^T, then summing the realized cumulative playtime. The solid line in Figure 2 shows the ensuing average realized playtimes for consumers if they had full information. On average, users in our sample could have achieved half of their status quo playtime with a very small share – 7.4 percent – of status quo expenditure. As the first vertical line indicates, users could have achieved the vast majority (90 percent) of status quo playtime with 40.3 percent of initial expenditure, or at a 59.7 percent discount. Hence, the last 10 percent of playtime costs consumers an average of roughly 12 times more, per hour, than the first 90 percent.

How close to full-information savings can users achieve by following sophisticated predictions? A consumer relying on sophisticated predictions h_{ij}^P could maximize expected playtime by purchasing games in ascending order of price per predicted hour, or p_j/h_{ij}^P. The dashed line in Figure 2 shows the maximum realized playtime (h_{ij}^T) achieved from reliance on these predictions. Sophisticated predictions achieve 90 percent of status quo playtime with a 24.8 percent discount, about 40 percent of the discount allowed by full information. It is of course conceivable that consumers value the time spent playing marginal games more highly than their prices per hour, so Figure 2 is merely suggestive of regret at this point.

5.2 Pre-purchase information, expenditure, and hours

The full effect of information depends not only on the regrettable purchases avoided but also on the beneficial purchase opportunities surfaced by better information. We calculate the overall effect of information, relative to the status quo, by comparing average hours of realized playtime delivered by status quo choices against three relevant alternatives that status quo expenditure could produce: random game choices, choices made with full information, and choices that follow sophisticated predictions.

Panel A of Figure 3 summarizes the resulting calculations using both owned and unowned games. First, randomly chosen bundles exhausting status quo budgets deliver on average 27.6 percent fewer hours of playtime than status quo choices, which implies that, on average, consumers have some useful information in the status quo. Second, a consumer armed with full information prior to purchase could on average nearly double status quo playtime (a 94.2 percent increase). Third, a consumer heeding a sophisticated predictions would on average achieve a 36.0 percent increase over status quo playtime, a little over a third of the effect of full information.

Panel B of Figure 3 depicts the descriptive results in expenditure rather than hours terms. The better the information the consumer has, the less costly is the achievement of status quo hours. The figure points to large welfare benefits of better information relative to the status quo. Full information allows just 19 percent of status quo expenditure – a $400 reduction in spending relative to the status quo. Hence, the average welfare benefit of full information to consumers is at least $400. Sophisticated prediction, analogously, allows users to benefit by at least $175, on average. Of course, the actual welfare effect of better information arises not only from a cost reduction for the achievement of status quo hours but also from users' informed choice of how many hours to purchase.

5.3 Measurement assumptions and playtime

We have made various measurement assumptions above that may drive the large estimated impacts of full information. These include assumptions about 1) the realized playtimes for both owned and unowned games, and 2) the measurement of prices and therefore status quo expenditure. We address each of these in turn below.

Playtime assumptions First, our main approach to measuring true playtime adds errors to predictions for unowned games that are sized to match the realized prediction errors for owned games. It is possible, however, that users' decisions not to purchase certain games reflect their knowledge that those games would deliver lower playtime. Then, our main approach would overstate the realization errors, and playtime, for unowned games. We explore this possibility by subtracting a sequence of differentials from our measures of true hours for non-owned games, in Panel A of Figure 4. The leftmost dots show the baseline results. The horizontal axis shows the proportionate reduction in the realized playtime values for unowned games. The larger the reduction, the smaller the effect of information on realized playtime. It is difficult to know what is a plausible upper bound, but if we shaded playtime for unowned games by 50 percent, full information would still allow status quo expenditure to produce a 45 percent increase over status quo playtime. At that level of shading, the playtime made feasible by sophisticated predictions would fall just six percent above status quo levels.

Second, we verify that we obtain similar estimates of the effect of additional information using four alternative ways of estimating true playtime for both owned and unowned games.

We obtain errors in two ways: We obtain random errors from the empirical distribution of deviations between true and predicted playtime; we also add parametric errors from a normal approximation to the error distribution. We use these two errors in two ways: We add these respective errors randomly to predictions only for unowned games; and we also estimate our true hours measures for the entire sample as the predictions (h^P_{ij}) plus these errors.

Using all four approaches, full information on average raises hours by between 92 and 128 percent, while sophisticated prediction raises average hours by between 45 and 61 percent. As when using the main measures of true playtime, full information nearly doubles realized playtime, while sophisticated prediction achieves just under half of the full information gains.

Pricing and ownership assumptions First, our main analysis treats sample games as though they were purchased at their average prevailing prices, but Steam sometimes makes games available at a discount or even free of charge. If users obtained games without payment, then our main analysis would overstate both users' status quo expenditure and the benefit that full information expenditure reallocation could deliver. We explore the robustness of our result to discounted games by recalculating the playtime gains available with full information and sophisticated predictions when we assume that a share T of each consumer's least-played games was obtained for free rather than at their average price. Reclassifying a purchased game as free reduces both our measured status quo expenditure and the scope for reallocation to raise possible playtime. Treating the least-played games as free, rather than as regrettable purchases, gives a lower bound of information effects at any share T and makes our robustness check conservative.

Panel B of Figure 4 illustrates playtime achieved when varying shares of owned games are assumed to have been free. At one extreme – the baseline case – no games are assumed to be free. Then full information reallocation of status quo expenditure raises playtime by the baseline 94 percent (the left dot on the solid line). At the other extreme, if all games were free, then there would be no status quo expenditure and therefore no additional playtime with full information. The ratio would be one, reflecting a zero percent increase in hours. The increase falls between the extremes for intermediate values of T. If 20 percent of users' games had been free, then full information would raise playtime by 80 percent. As the share of games obtained without charge rises, the full information benefit falls; but the full information increase in playtime remains above 50 percent with up to 80 percent of games obtained without charge. The same exercise shows that sophisticated predictions also raise playtime at status quo expenditure with up to 70 percent of games obtained free of charge. We conclude that free games are not likely to explain our basic results.

Second, users are eligible to return games they played less than two hours, and we know that 5 to 8 percent of games are returned. We rationalize this fact in our main analysis by assuming that games played less than 23 minutes are returned. To verify that this cutoff is not driving our results, we recalculate the additional playtime that additional information can deliver for a range of thresholds from 0 to 120 minutes. The proportionate effect of full information on the playtime that status quo expenditure buys runs from 1.96 to 1.90 as the threshold rises from 0 to 120 minutes. The proportionate playtime effect of sophisticated predictions runs from 1.39 to 1.29. Results are essentially unchanged relative to the main analysis.

5.4 Modeling bundles with aggregate hours

Our basic modeling approach assumes that consumers value bundles according to the aggregate hours they deliver. This, in turn, implies that the marginal utility of hours is equal across games. The data on game purchases can be used to explore the reasonableness of this assumption. The idea is simple: if people value an hour of play similarly across games, then they will be more willing to buy games delivering more hours, all things equal. To explore this, we aggregate the data to the game level. Define s_j as the share of our users owning game j and H_j as a measure of the average hours delivered by game j.[20] We postulate that consumer i's utility for game j depends on a function of the average hours the game delivers, its price, a game-specific unobservable ξ_j, and an extreme value error; and we estimate this model as a simple logit, using levels and logs of both hours measures, via

$$\ln({}^{s_j}/_{1-s_j}) = \beta_0 + \beta_j f(H_j) + \alpha p_j + \xi_j. \tag{3}$$

For all four regressions, the estimates yield positive and significant coefficients on hours and negative and significant price coefficients. The left panel of Figure 5 depicts the relationship between s_j and $\ln(H_j)$ using observed hours of playtime, and a positive relationship is clearly evident. This supports the idea that users purchase games for the hours that they deliver, which itself lends support to aggregation of hours across games.

Yet, the points in the left panel of Figure 5 are not precisely on a line. Although this is partly because the figure does not account for prices, unobserved game ownership tendencies (ξ_j) differ additionally for a variety of possible reasons, including different marginal utilities of hours across games. We explore this by loading all of the variation in ξ_j into game-specific hours coefficients in Equation (3). This gives game-specific weights ω_j which solve $\beta \ln(H_j) + \xi_j = \beta \ln(\omega_j H_j)$, so $\omega_j = \left(e^{\xi_j}\right)^{(1/\beta)}$. We use degrees of these weights to create a range of adjusted hours measures. User i's weighted hours measure is then:

$$H_i^* = \sum_{j \in J} \left(\lambda\omega_j + (1-\lambda) \times 1\right) h_{ij} \mathbb{1}_{ij},$$

where λ ranges from 0 (unweighted estimates) to 1 (full weighting).

The right panel of Figure 5 shows weighted estimates of the proportional gains in playtime at status quo expenditure, for full information and prediction, and various values of λ. The

[20] We use two measures, the average of observed hours among owners and the average measure of true hours across all owners.

leftmost dots (at $\lambda = 0$) reproduce the baseline results from Figure 3 Panel A. Two things are clear. First, even when we attribute all cross-game variance in ownership to different marginal utilities (when $\lambda = 1$), the gain from full information remains at 60 percent. Second, sophisticated predictions achieve about 40 percent of the full information gains for all values of λ. We conclude, first, that modeling the bundle through the hours it delivers is reasonable and, second, that the assumption implicit in our use of aggregate hours in the utility function does not drive our main finding of substantial opportunities forgone in status quo consumption.

6 Structural model

The descriptive results suggest substantial welfare forgone in status quo consumption, relative to full information, and that sophisticated predictions could allow consumers to attain an appreciable share of the full information benefit. Yet, the calculations are rough in that they neither allow endogenous selection of playtime, nor do they quantify the welfare benefit of information in dollar terms. To analyze the welfare effects of predictions in a theory-consistent way, we need two major ingredients. First, we need to calculate status quo, full information, and prediction-informed budget sets, which we obtain using the framework from Section 3. We do this in Section 6.1. Second, we need an estimated utility function for selecting the utility-maximizing points on the respective budget constraints; and Section 6.2 discusses Cobb Douglas and logit utility function implementations based on aggregate hours.

6.1 Information and the expected budget constraint

The budget constraint that each consumer faces depends on their information, which is embodied in the model through the parameter κ_i in Equation (2). In particular, we model each consumer's rank ordering of games as depending on an index which is a convex combination of the random and full information ranks, $I(\kappa_i) = \kappa_i r^T + (1 - |\kappa_i|)r^\epsilon$, where $\kappa \in [-1, 1]$.

We simulate the index based on 50 draws of the random ranking r^ϵ for each user and each value of $\kappa \in [-1, -0.9, \dots, 0.9, 1]$. For each user and κ, we compute the average hours of playtime that each user's status quo expenditure would deliver, $H_i(P_i; \kappa)$, and we choose the κ_i that minimizes the distance between the average simulated playtime $H_i(P_i; \kappa)$ and the status quo playtime H_i.

Each draw gives a user's budget constraint. We calculate the slope of the status quo budget constraint for each user as the average game price over the average hours delivered for the game purchase where simulated cumulative expenditure is closest to status quo expenditure. The resulting user-specific slopes of the status quo budget constraints ($p_H(\kappa)$) imply that an additional hour costs an average of \$0.56 (median of \$0.42) at users' status quo bundles. This varies between \$0.26 at the 25th percentile and \$0.69 at the 75th.

6.2 Cobb Douglas and logit utility functions

We first estimate welfare effects using a Cobb-Douglas calibration. We observe an initial hours choice for each user, which (suppressing i subscripts) we now denote by H_0; and the κ derivation above delivers the status quo budget constraint slope at H_0. This allows us to infer the initial levels of "all other goods," or AOG level A_0 and therefore to calculate initial utility. Given the hours expenditure share a (which is informed by expenditure data), we have the utility function $U = H^a A^{1-a}$; and user utility maximization arises where their MRS equals the slope of the budget constraint, $p_H(\kappa)$. Rearranging terms, we solve

$$A_0 = \frac{1-a}{a} H_0 p_H(\kappa)$$

for each user. Status quo utility is then $U_0 = H_0^a A_0^{1-a}$.

In order to generate the prediction-informed and full information budget constraints, we need the point where the status quo, full information, and prediction-informed budget constraints meet the AOG axis (the user's income level). We calculate this as $I = A_0 + P_0$. Given I, we create the full information and prediction-informed budget constraints and calculate the choices for each user as the utility-maximizing points along these budget constraints. The value of information, relative to status quo choices, is the amount of money the user would need to forgo from an informed state to bring their utility to the status quo level.

While the Cobb Douglas setup allows for utility maximization and – by construction – fits status quo decisions perfectly, these estimates have the shortcoming of imposing the utility function parameters based on aggregate spending shares, rather than using the data to estimate the way in which consumers trade off hours for dollars. We do this with a logit framework. Like Crawford and Yurukoglu (2012), we model the utility of each bundle as a function of bundle characteristics and an extreme value error. Specifically, the utility that user i derives from their chosen bundle with playtime H_i and expenditure P_i is given by

$$u_{ib} = \beta \ln(H_i) - \alpha P_i + \mu_i + \epsilon_{ib}, \tag{4}$$

where ϵ_{ib} is an extreme value error, and b refers to a game bundle available to user i. We employ the log specification to reflect the diminishing marginal utility of game playtime, and we include user fixed effects (μ_i) to account for variation in overall utility of hours played.

We estimate the parameters of the utility function in Equation (4) by noting that consumer i chose their owned bundle and not other bundles. But what other bundles should we understand them to have rejected? Define $\bar{h}$ as the average hours delivered by a game. Then we model the user as preferring the status quo choice to either $\bar{h}$ additional, or fewer, hours of playtime at the user's status quo price per hour $p_H(\kappa)$. That is, the user chose the observed bundle (H_i, P_i) over two alternative bundles described by the hours and price tuples $(H_i + \bar{h}, P_i + p_H(\kappa)\bar{h})$ and $(H_i - \bar{h}, P_i - p_H(\kappa)\bar{h})$.

The top panel of Table 3 (Model 1) reports estimation results from this specification, which are our main estimates. The coefficients indicate that consumers value both additional hours and money. Given that the mean cumulative playtime is 2,166.4 hours, the implied average value of an additional hour of play is \$0.60, which is similar to the average cost per hour above from the budget constraints (\$0.56).[21]

In the logit welfare simulations below, we use the estimated coefficients to find the utility-maximizing bundles under full information and sophisticated predictions. Both the coefficients and the chosen bundles also inform the CS effects of information. For example, the effect of full information on a particular user's consumer surplus is given by

$$\Delta CS_i = \frac{1}{\alpha} \ln \left(1 + e^{\beta \ln(H_i^{FI}) - \alpha P_i^{FI}}\right) - \frac{1}{\alpha} \ln \left(1 + e^{\beta \ln(H_i^{SQ}) - \alpha P_i^{SQ}}\right), \tag{5}$$

where the superscripts FI_i and SQ_i denote the hours and expenditures from the bundles chosen with full information and in the status quo, respectively.

6.3 Utility function extensions

Before turning to welfare estimates, we explore two extensions of the baseline (Model 1) logit specification, involving users of differing tenure on the platform, as well as separate terms for the cumulative hours of different game types.

First, our baseline approach embodies diminishing marginal utility of hours of playtime during the same time interval, but users in our sample have been on the platform for different amounts of time, and this could distort our utility function parameter estimation. To address this, we re-estimate the model separately for groups of users who joined in each year, from

[21] We obtain 0.60 from $(\hat{\beta}/\bar{H})/\hat{\alpha}$.

2003 to 2016. As the middle panel of Table 3 (Model 2) shows, we find positive log hours parameters and negative price parameters for all cohorts; and we cannot reject the joint hypotheses that all price, and log hours, parameters are equal across user join years. We conclude that our sample's inclusion of users with different tenures on the platform does not distort the baseline parameter estimates. We also report welfare estimates from this flexible specification below.

Second, in the foregoing exercises, utility depends on the sum of hours from all games. To allow for the possibility that different games types enter the utility function separately, we divide games into two types: action games and others. Among the 100 games in our dataset, 71 include an "action" tag in their description and 29 do not. We adjust the estimation equation in (4) to allow hours of different game types to affect utility differently:

$$u_{ib} = \beta^{A} \ln(H_i^{A}) + \beta^{NA} \ln(H_i^{NA}) - \alpha P_i + \mu_i + \epsilon_{ib}, \tag{6}$$

where the A superscript refers to action games, NA refers to non-action games, and b refers to the bundle of games chosen by user i. As above, we estimate the equation as a fixed effects logit. Here, the observed choice is modeled as preferred to four non-chosen bundles with one more, or one fewer, action or non-action game.

The bottom panel of Table 3 (Model 3) reports the estimated parameters. While the (common across game types) consumer disutility of expenditure is roughly similar to that in the one-type case, consumer utility of hours played varies significantly across game types: Users value additional playtime for action games more than for non-action games. Despite these differences, the marginal utilities of additional hours – at the means of action and non-action games played – are similar.[22] We show below that the welfare effects from this model are similar to the baseline effects.

[22] We exclude one user, for whom measured non-action playtime is under 5 minutes, from this calculation.

7 Welfare effects of information

Given the model estimates, we compare status quo choices to alternatives with different information to measure the welfare effect of information. Sections 7.1 presents estimates of the benefits from avoiding regrettable purchases, relying only on owned games. Section 7.2 presents estimates of the overall welfare effects, including the effects of being made aware of otherwise-unknown products. We explore the variation in these effects across users and games in Section 7.3. Finally, Section 7.4 compares the dollar values of the benefits achievable with progressively more sophisticated predictions.

7.1 Gains from avoiding regret

While the full welfare gains we calculate arise from better information about both purchased and non-purchased games, we begin with a simpler calculation of the gains from avoiding purchases that users regretted. We solve the utility models against budget constraints that include only the games purchased in the status quo.

Table 4 presents the no-regret results using both the logit and Cobb Douglas approaches. The top panel shows effects from full information. By construction, expenditure, games purchased, and hours decrease. With full information about already-owned games, users would buy an average of 11.0 (11.9 for Cobb Douglas) of the 33.6 games they had purchased in the status quo, reducing their expenditure by \$354.9 (\$339.1), or by about two thirds.[23] Because they are just eliminating games, playtime must fall; but the average hours reduction is only about 11.5 (13.8) percent. Given the large expenditure reductions and relatively small playtime reductions, consumer surplus rises by an average of \$261.2 (\$245.3), or by nearly half of status quo expenditure.

The bottom panel of Table 4 shows the effects of heeded sophisticated predictions on

[23] We report Cobb Douglas estimates based on $a = 0.05$. Results are nearly identical with $a = 0.01$ or $a = 0.005$.

regrettable status quo purchases. Game purchases fall from 33.6 to 17.7 (19.5), and expenditure falls from $508.1 to $240.6 ($272.6). Playtime falls proportionately more than with full information, however, by 25.6 (21.1) percent, suggesting that games eliminated under sophisticated predictions are better than those eliminated under full information. Still, average CS rises by $73.7 ($53.9), or by about a tenth of status quo expenditure. CS rises for 76.4 percent of users.

7.2 Overall welfare effects of information

The overall effect of better pre-purchase information arises not only from avoiding regret but also from being alerted to previously-unknown opportunities. The top panel of Table 5 reports overall welfare effects of full information. While avoiding regret by itself reduces both purchases and hours played, full information about all games raises hours played substantially. Hours rise by roughly three fifths above the status quo level. At the same time, expenditures (and games purchased) decline by about a half. Full information raises consumer surplus by an average of $682.4 ($626.4), or by 134.3 (123.2) percent of the $508.1 in status quo expenditures. These estimates of the change in CS are roughly 2.5 times the effects from eliminating regret alone. Hence, full information would allow substantially higher utility with much less expenditure, relative to the status quo; and much of the overall gain stems from the purchase of otherwise-unknown games.

The second panel of Table 5 shows the overall effects of sophisticated prediction, if heeded. Relative to the status quo, consumers would reduce their game expenditures by 38 (30) percent – and their numbers of games by 28 (19) percent – while increasing average playtime by about ten percent, relative to the status quo. Access to the formerly-unknown opportunities revealed by the predictions allows average CS to rise by $292.5 (242.9), roughly half of status quo expenditure. Hence, sophisticated predictions produce between a third and two fifths of full information benefits.

Despite our failure to reject constant parameters across users with different Steam platform tenure, we also estimate the average changes in CS from both full information and sophisticated predictions, by join year, using the separate parameters by year. Figure 6 reports these results, and the average full information change in CS is roughly \$660 across join years, which is similar to the baseline estimate of \$682.4. The average change from sophisticated predictions is about \$290, nearly the same as the baseline estimate of \$292.5.

The welfare effects of full information using the two game types are also very similar to the estimates above. Hours played rise to about 3,205.5 rather the 3,461.8 in the baseline logit. The number of games owned falls to 12.7 rather than 14.4, and the change in CS is \$554.7 rather than \$682.4.

7.3 Distributional effects of information on users and game sellers

Users. Not only does full information deliver changes in CS and expenditure, but these outcomes also vary substantially across users. The inter-quartile range in ΔCS runs from \$489.4 to \$811.6, and the inter-quartile range for expenditure reduction runs from \$153.4 to \$429.1. The value of sophisticated predictions also varies substantially across users: the inter-quartile range in ΔCS runs from \$123.4 to \$447.5. The predictions are not guaranteed to improve on status quo information, but sophisticated predictions raise CS for 88.5 percent of users.

Which sorts of users experience larger changes in CS or expenditure? In our modeling framework, changes in expenditure and CS depend on two factors. First, consumers can be differently informed in the status quo. Users with poor information – low or negative κ_i and therefore less bowed-out budget constraints – get larger benefits from additional information, all else constant. Second, users for whom the price of hours varies substantially across games (i.e., with a higher variance of $^{h^T_{ij}}/_{p_j}$) have more to gain from additional information.

Empirically, both matter; but the variation in κ gives rise to much more of the variation in ΔCS and changed expenditure. We run the regression $\Delta CS_i = \nu_0 + \nu_1 \kappa_i + \nu_2 \sigma_i \left({}^{h_{ij}^T}/{}_{p_j} \right) + \varepsilon_i$, where the σ term is the user-specific standard error of h/p. We estimate $\hat{\nu_1}$ = -1053.0 (se = 2.53) and $\hat{\nu_2} = 0.34$ (0.01). These coefficients imply that a one-standard deviation increase in κ changes ΔCS by 6.6 times the change for a one-standard deviation increase in our measure of the variation in h/p.

The source of the welfare gain from better information varies across users with their changes in expenditure. Users who decrease expenditure when fully informed benefit from avoiding regrettable purchases, while those who spend more when fully informed benefit from being alerted to enjoyable games. This gives rise to "U-shaped" relationships between changes in expenditure and CS in Figure 7. The left panel shows full information ΔCS against expenditure, while the right panel shows prediction-informed ΔCS. The blue smoothed lines show the changes in ΔCS arising only from the elimination of regrettable purchases. A comparison of the black and blue lines shows that predictions mainly allow avoidance of regret, while full information alerts users to games they otherwise would not have bought.

Sellers. While consumers unambiguously gain from full information, this comes at sellers' expense. Full information reduces expenditure substantially – by 26 percent – overall, so games sales fall. Figure 8 plots game sales quantities in the informed state (y-axis) against status quo sales (x-axis), and full information reduces quantities sold for over 90 percent of sample games. Sales fall for both popular and less popular sample games, so information does not appear to change the concentration of sales. Overall, better information would bring about a large transfer from sellers to buyers. Because the full information ΔCS is more than twice as large as the reduction in expenditure, total surplus would increase under full information. Of course, revenue effects would differ if prices were not held constant across information environments; and patterns of demand changed by better pre-purchase information could affect prices and revenue.

7.4 The value of better prediction technology

The effects of personalized predictions calculated above depend on the predictive accuracy of the chosen model. How do the welfare effects of our preferred prediction model – with 100-factor predictions – compare with alternatives? RMSE is a natural statistical way to evaluate prediction approaches, but RMSE does not attach a dollar value to accuracy. We explore this in Figure 9, which reports the changes in CS relative to the status quo arising from a sequence of predictions. Predictions based on user averages deliver \$21.4 less in average CS than the status quo, while matrix factorization with 10 latent factors raises CS by \$203.7 above status quo levels. Predictions based on game averages raise CS by \$240.5 above the status quo, and our 100 latent factor model using half of the data raises per capita CS similarly, by \$247.8. Matrix factorization with 50 latent factors, and using all of the data, raises CS by \$267.1 above the status quo, similar to the effect of the random forest prediction. Finally, our preferred prediction approach raises CS by \$292.5 relative to the status quo, delivering 43 percent of the advantage of full information. Quantification of the value of better prediction may be useful both for guiding social investment in prediction technology and for understanding the costs of privacy.

8 Conclusion

Differentiated products can deliver substantial value to heterogeneous consumers, but only if people know which products to purchase. Lack of post-purchase usage data has diverted attention from the possible problems of regret and missed opportunities in differentiated product choices. Using novel data on post-purchase usage of video games, we document that status quo purchases deliver benefits that fall far short of what full information would allow. Descriptive analysis shows that full information would allow consumers to purchase nearly double the hours of playtime with status quo expenditures. Consumers following sophisticated, personalized predictions could achieve roughly 40 percent of this additional usage time.

To measure the welfare effects, we develop an explicit measurement framework in two parts. First, we develop a model of how information affects choice sets; and second, we create a tractable model of consumer choices of product bundles. Using a Cobb Douglas calibration and logit estimates, we draw two major conclusions. First, status quo consumption forgoes a great deal of potential benefit. In our setting, full information could raise CS by about 130 percent of status quo expenditure while cutting expenditure in half. Second, sophisticated prediction approaches, if heeded, allow recovery of about 40 percent of this untapped potential welfare benefit. We also document that more information – about either consumers or products – can allow for predictions that yield more of the welfare benefit.

The problems we study might arise in a variety of contexts in which heterogeneous consumers choose among differentiated products. Given suitable data, documenting the shortcomings of status quo consumption for other contexts is a fruitful area for further study.

References

Adomavicius, G. and J. Zhang (2012): “Stability of recommendation algorithms,” *ACM Transactions on Information Systems*, 30, 1–31.

Aguiar, L. and J. Waldfogel (2018): “Quality predictability and the welfare benefits from new products: Evidence from the digitization of recorded music,” *Journal of Political Economy*, 126, 492–524.

Aguiar, M., M. Bils, K. K. Charles, and E. Hurst (2021): “Leisure luxuries and the labor supply of young men,” *Journal of Political Economy*, 129, 337–382.

Allcott, H. (2013): “The welfare effects of misperceived product costs: Data and calibrations from the automobile market,” *American Economic Journal: Economic Policy*, 5, 30–66.

Anderson, C. (2007): *The long tail: How endless choice is creating unlimited demand*, Random House.

Anderson, S. P., A. De Palma, and Y. Nesterov (1995): “Oligopolistic competition and the optimal provision of products,” *Econometrica: Journal of the Econometric Society*, 1281–1301.

Bergstra, J. and Y. Bengio (2012): “Random search for hyper-parameter optimization.” *Journal of Machine Learning Research*, 13.

Bobadilla, J., F. Ortega, A. Hernando, and A. Gutiérrez (2013): “Recommender systems survey,” *Knowledge-Based Systems*, 46, 109–132.

Breiman, L. (2001): “Random forests,” *Machine Learning*, 45, 5–32.

Brynjolfsson, E., Y. Hu, and M. D. Smith (2003): “Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers,” *Management science*, 49, 1580–1596.

Chevalier, J. A. and D. Mayzlin (2006): “The effect of word of mouth on sales: Online book reviews,” *Journal of marketing research*, 43, 345–354.

Chu, C. S., P. Leslie, and A. Sorensen (2011): “Bundle-size pricing as an approximation to mixed bundling,” *The American Economic Review*, 263–303.

Crawford, G. S. and A. Yurukoglu (2012): “The welfare effects of bundling in multichannel television markets,” *American Economic Review*, 102, 643–685.

Dixit, A. K. and J. E. Stiglitz (1977): “Monopolistic competition and optimum product diversity,” *The American economic review*, 67, 297–308.

Donnelly, R., A. Kanodia, and I. Morozov (2023): “Welfare effects of personalized rankings,” *Marketing Science*.

Einav, L., B. Klopack, and N. Mahoney (2023): “Selling Subscriptions,” Tech. rep., National Bureau of Economic Research.

Funk, S. (2006): “Netflix update: Try this at home,” .

Jannach, D., P. Resnick, A. Tuzhilin, and M. Zanker (2016): “Recommender systems—beyond matrix completion,” *Communications of the ACM*, 59, 94–102.

Kaye, A. (2023): “The Personalization Paradox: Welfare Effects of Personalized Recommendations in Two-Sided Digital Markets,” *https://apkaye.github.io/aaronpkaye.website/Kaye_Aaron_JMP.pdf*.

Koren, Y. (2009): “Collaborative filtering with temporal dynamics,” in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 447–456.

Koren, Y., R. Bell, and C. Volinsky (2009): “Matrix factorization techniques for recommender systems,” *Computer*, 42, 30–37.

Koren, Y., S. Rendle, and R. Bell (2021): “Advances in collaborative filtering,” *Recommender systems handbook*, 91–142.

Lee, D. and K. Hosanagar (2021): “How do product attributes and reviews moderate the impact of recommender systems through purchase stages?” *Management Science*, 67, 524–546.

Lee, R. S. (2013): “Vertical integration and exclusivity in platform and two-sided markets,” *American Economic Review*, 103, 2960–3000.

Mankiw, N. G. and M. D. Whinston (1986): “Free entry and social inefficiency,” *The RAND Journal of Economics*, 48–58.

Miravete, E. J. (2003): “Choosing the wrong calling plan? Ignorance and learning,” *American Economic Review*, 93, 297–310.

Nelson, P. (1970): “Information and consumer behavior,” *Journal of political economy*, 78, 311–329.

Quan, T. W. and K. R. Williams (2018): “Product variety, across-market demand heterogeneity, and the value of online retail,” *The RAND Journal of Economics*, 49, 877–913.

Reimers, I. and J. Waldfogel (2021): “Digitization and pre-purchase information: the causal and welfare impacts of reviews and crowd ratings,” *American Economic Review*, 111, 1944–1971.

Reinstein, D. A. and C. M. Snyder (2005): "The influence of expert reviews on consumer demand for experience goods: A case study of movie critics," *The journal of industrial economics*, 53, 27–51.

Rendle, S., W. Krichene, L. Zhang, and J. Anderson (2020): "Neural collaborative filtering vs. matrix factorization revisited," in *Proceedings of the 14th ACM Conference on Recommender Systems*, 240–248.

Skelton, A. C. and J. M. Allwood (2017): "Questioning demand: a study of regretted purchases in Great Britain," *Ecological Economics*, 131, 499–509.

Spence, M. (1976): "Product differentiation and welfare," *The American Economic Review*, 66, 407–414.

Sun, T., Z. Yuan, C. Li, K. Zhang, and J. Xu (2024): "The value of personal data in internet commerce: A high-stakes field experiment on data regulation policy," *Management Science*, 70, 2645–2660.

Waldfogel, J. (2007): *The tyranny of the market: Why you can't always get what you want*, Harvard University Press.

Ward, M. R. (2010): "Video games and adolescent fighting," *The Journal of Law and Economics*, 53, 611–628.

Wu, R., Y. Huang, and N. Li (2023): "Platform Information Design and Competitive Price Targeting," *Available at SSRN*.

9 Tables and figures

Table 1: Summary statistics

	by ownership status		total
	owned games	non-owned	all games
Price	15.10	15.28	15.22
	(9.36)	(9.78)	(9.64)
True ln(minutes)	6.65	5.80	6.08
	(1.75)	(2.19)	(2.09)
Predicted ln(minutes)	6.40	5.93	6.09
	(1.12)	(1.13)	(1.15)
N	1,682,214	3,317,786	5,000,000

Note: Averages and standard deviations for expenditure and playtime for owned and unowned games (columns 1 and 2), as well as for all games in the dataset (column 3). All summary statistics are at the user-game level, and playtime measures are in natural logs of minutes. We treat games played less than 23 minutes as unowned.

Table 2: RMSE for different prediction models

	RMSE			
Approach	**Train**	**Test**	**Validation**	**Details**
Collab. Filtering	**0.286**	**0.589**	**0.607**	**Our model,** $k = 100$
	0.314	0.596	0.611	$k = 50$
	0.467	0.621	0.639	$k = 10$
	0.524	0.622	0.647	$k = 5$
Game & user FEs	0.589	0.611	0.627	
Game & user char's	0.616	0.623	0.640	Random Forest (nonlin.+interactions)
Game & user char's	0.625	0.630	0.645	All characteristics from below
Game FEs	0.629	0.634	0.648	
Game char's	0.629	0.634	0.648	6 attributes & 95 tag dummies
Game char's	0.684	0.688	0.701	6 attributes
User char's	0.703	0.708	0.720	7 attributes & 117 country dummies
User char's	0.704	0.708	0.721	7 attributes
Global average	0.706	0.711	0.722	

Note: Prediction error (RMSE) for training, test, and validation sets for a data subset of 39,130 users for whom user-level characteristics are available and 100 games with at least partial game-level level data. The models are estimated with OLS unless stated otherwise. Game and user attributes used in the models are described in detail in Appendix Section A, and missing observations were imputed with means.

Table 3: Logit estimates

	ln(H)		Price	
	coef.	std. err.	coef.	std. err.
Model 1:				
baseline	1.266***	(0.104)	–0.0020***	(0.0002)
Model 2:				
2003	1.181*	(0.617)	–0.0017	(0.0011)
2004	1.309***	(0.403)	–0.0021***	(0.0007)
2005	1.054*	(0.581)	–0.0016	(0.0010)
2006	1.433**	(0.597)	–0.0021**	(0.0010)
2007	1.349**	(0.542)	–0.0021**	(0.0009)
2008	1.334**	(0.571)	–0.0020**	(0.0010)
2009	1.09***	(0.413)	–0.0016**	(0.0007)
2010	1.059***	(0.397)	–0.0015**	(0.0007)
2011	1.297***	(0.477)	–0.0018**	(0.0008)
2012	1.396***	(0.418)	–0.0021***	(0.0007)
2013	1.528***	(0.405)	–0.0025***	(0.0008)
2014	1.428***	(0.497)	–0.0022**	(0.0009)
2015	1.416***	(0.46)	–0.0025**	(0.0010)
2016	1.584***	(0.488)	–0.0030***	(0.0011)
other	1.189***	(0.183)	–0.0019***	(0.0004)
Model 3:				
action	1.169***	(0.056)	–0.0035***	(0.0001)
non-action	0.353***	(0.021)	–0.0035***	(0.0001)

Note: Conditional logit estimates of game demand at the user level. For each individual i we construct hours (H_i) and associated expenditure (P_i) for the owned bundle as well as non-chosen alternative bundles as described in the text. The first row (Model 1) reports coefficients for the baseline logit model, imposing the same utility of playtime across games. The middle panel (Model 2) reports coefficients for models estimated separately for users with different Steam join years. The bottom panel (Model 3) allows different coefficients for action and non-action games, with a common expenditure coefficient. The models are estimated on all 50,000 users. We treat games played less than 23 minutes as not owned.

Table 4: Welfare results – avoiding regret

	hours	# games	game expenditure	ΔCS
actual	2166.44	33.64	508.10	
Full information				
Logit	1867.94	10.95	153.18	261.15
Cobb Douglas	1917.95	11.91	169.00	245.34
Predictions				
Logit	1611.95	17.69	240.61	73.71
Cobb Douglas	1711.17	19.48	272.62	53.90

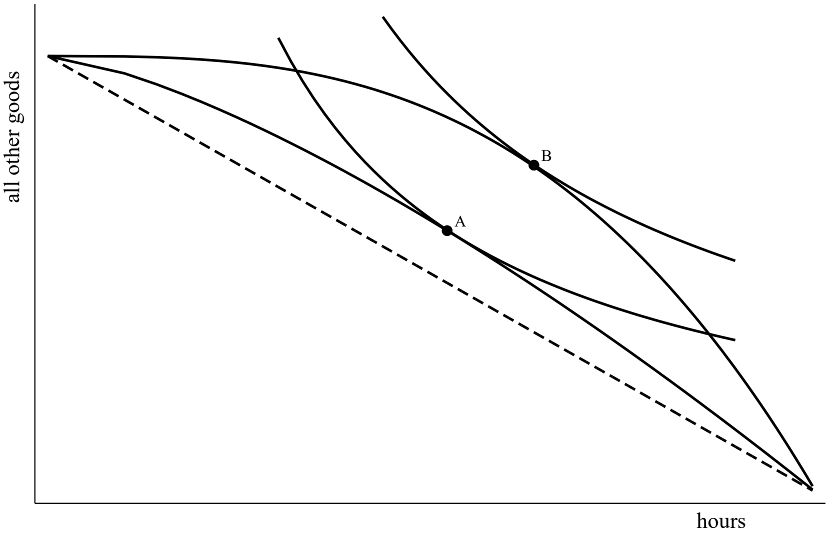
Note: Simulations of the effects of full information, and sophisticated, personalized prediction, on purchase and usage of, as well as consumer surplus from, games bought in the status quo. These simulations quantify the value of avoiding regrettable status quo purchases. The first row shows status quo values of hours, the number of games purchased, and game expenditure; and the remaining rows show these measures, as well as the change in consumer surplus, from various simulations. The next two rows show effects of full information using logit and Cobb Douglas models, respectively. The last two rows show logit and Cobb Douglas-based measures of the effects of heeding sophisticated predictions. All figures are per-person averages.

Table 5: Welfare results – overall effects

	hours	# games	game expenditure	ΔCS
actual	2166.44	33.64	508.10	
Full information				
Logit	3461.80	14.39	201.52	682.40
Cobb Douglas	3472.12	15.95	224.70	626.43
Predictions				
Logit	2360.99	24.35	312.53	292.49
Cobb Douglas	2433.40	27.28	360.47	242.86

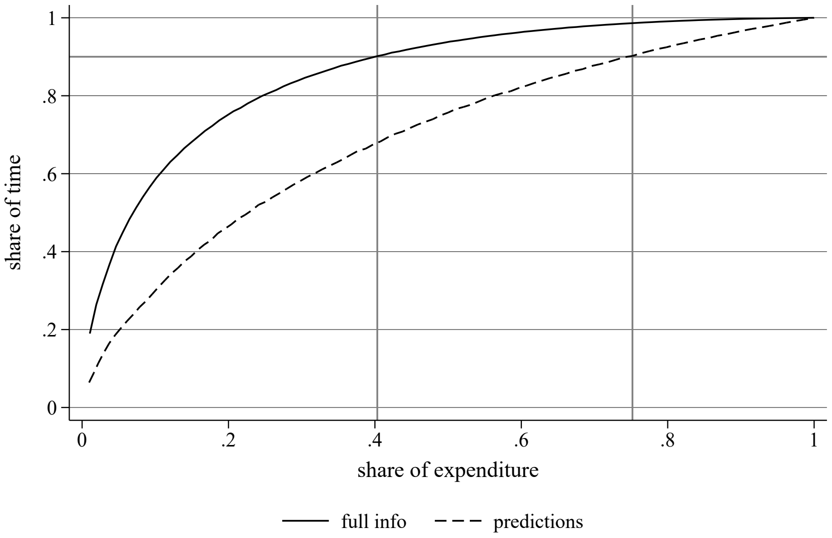
Note: Simulations of the effects of full information, and sophisticated, personalized prediction, on purchase and usage of, as well as consumer surplus from, all games in the dataset. These simulations quantify the combined value of both avoiding regrettable status quo purchases and finding otherwise-unknown games. The first row shows status quo values of hours, the number of games purchased, and game expenditure; and the remaining rows show these measures, as well as the change in consumer surplus, from various simulations. The middle two rows show effects of full information using logit and Cobb Douglas models, respectively. The last two rows show logit and Cobb Douglas-based measures of the effects of heeding sophisticated predictions. All figures are per-person averages.

Figure 1: Pre-purchase information, hours of playtime, and expenditure on games



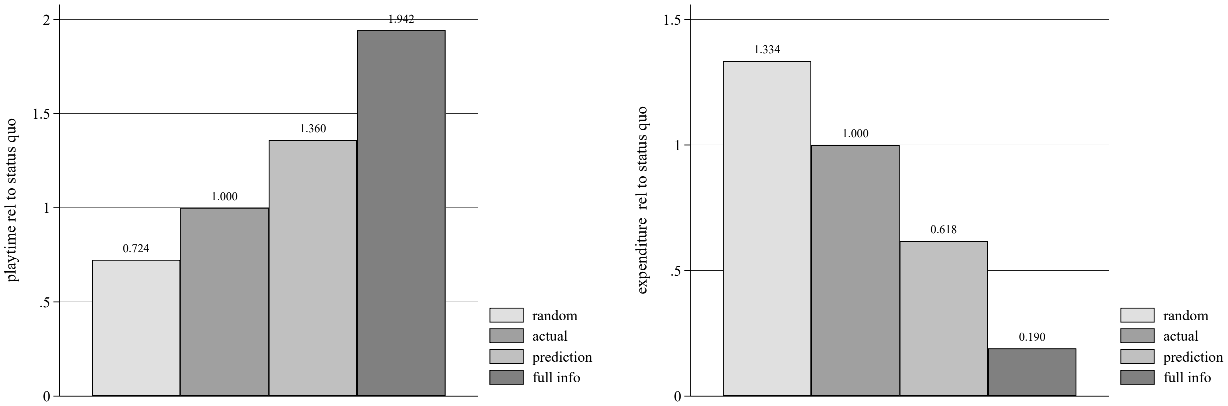
Notes: The figure depicts two budget constraints for hours of playtime (x-axis) vs all other goods (y-axis). The outer budget constraint reflects full information about games' playtime and prices, while the inner curved budget constraint reflects imperfect but better-than-random information on games' playtime in relation to their prices. A fully informed consumer maximizes utility by choosing point B, while the imperfectly informed consumer maximizes utility by choosing A. The less accurate the information the consumer has, the less "bowed out" the budget constraint. A consumer with no information about games would face the (dashed) straight-line budget constraint.

Figure 2: Potential for regret



Notes: This figure plots the share of status quo playtime delivered by shares of status quo expenditure on owned games. The solid line reflects full information: games are ordered by realized playtime per dollar spent. It shows that users could on average attain 90 percent of status quo playtime with just over 40 percent of status quo expenditure. The dashed line shows the effects of heeding sophisticated predictions: A user heeding such predictions could achieve 90 percent of status quo playtime with roughly 75 percent of status quo expenditure. The two vertical line illustrate these shares.

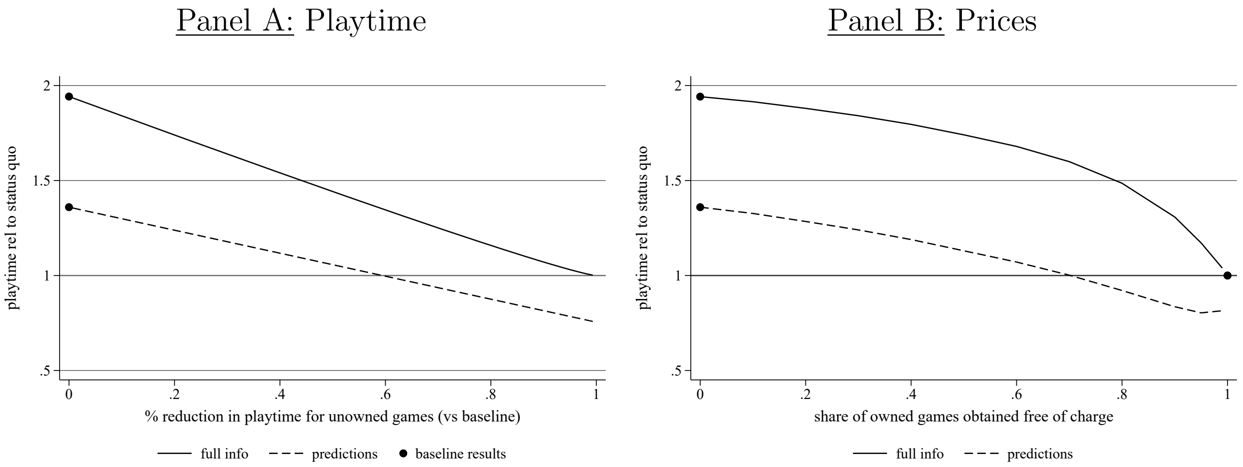
Figure 3: Potential for welfare gains


Panel A: playtime at s.q. expenditure

Panel B: expenditure at s.q. playtime

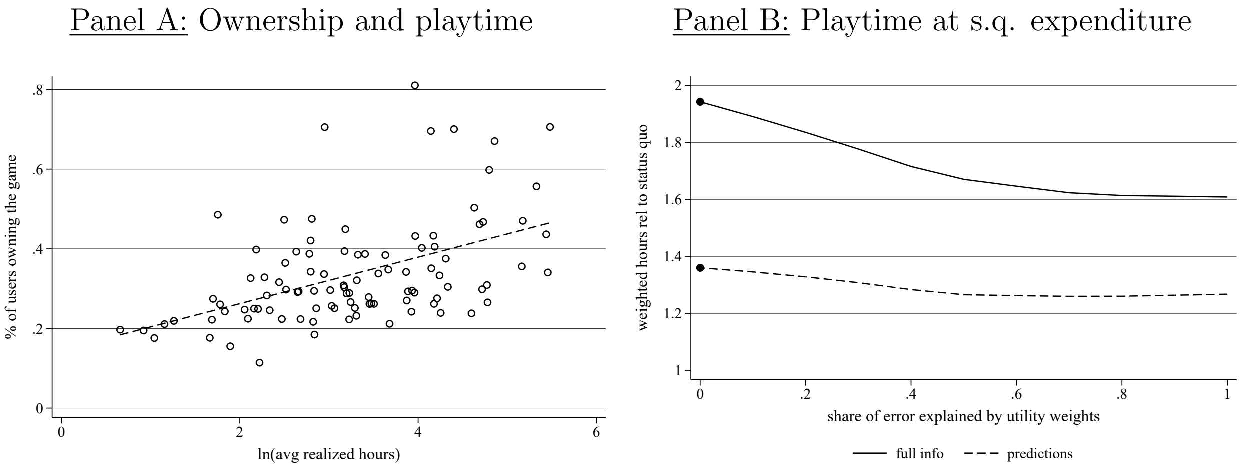
Notes: The bars in Panel A show the average cumulative playtime, relative to the status quo, that status quo expenditure could purchase under different information assumptions. For example, the leftmost bar indicates that random selection of games would deliver 72.4 percent of status quo playtime, while full information would allow the achievement of 94.2 percent more playtime. Panel B shows the results in terms of expenditure needed to achieve status quo playtime. For example, users buying random games would need to spend 33.4 percent more than status quo consumers for the same playtime, while fully informed users could attain status quo playtime with 19 percent of status quo expenditure.

Figure 4: Playtime at status quo expenditure for varying measurement assumptions



Notes: This figure plots the playtime that can be achieved at status quo expenditure, relative to status quo playtime, for varying assumptions about realized playtime for unowned games (panel A), and for varying assumptions about prices paid for owned games (panel B). The solid line in Panel A shows the playtime attainable with full information, relative to status quo playtime, for various degrees of shading of unowned games' playtime. For example, if unowned games delivered 40 percent less usage than owned games, then full information would allow 45 percent, rather than 94 percent, more playtime. The dashed line does the analogous calculation for prediction-informed choices. Panel B performs similar calculations based on the shares of owned games assumed to be obtained free rather than at their average prices. For example, if 80 percent of games had been free, then full information would allow 50 percent, rather than 94 percent, more playtime.

Figure 5: Robustness to varying utility weights across games

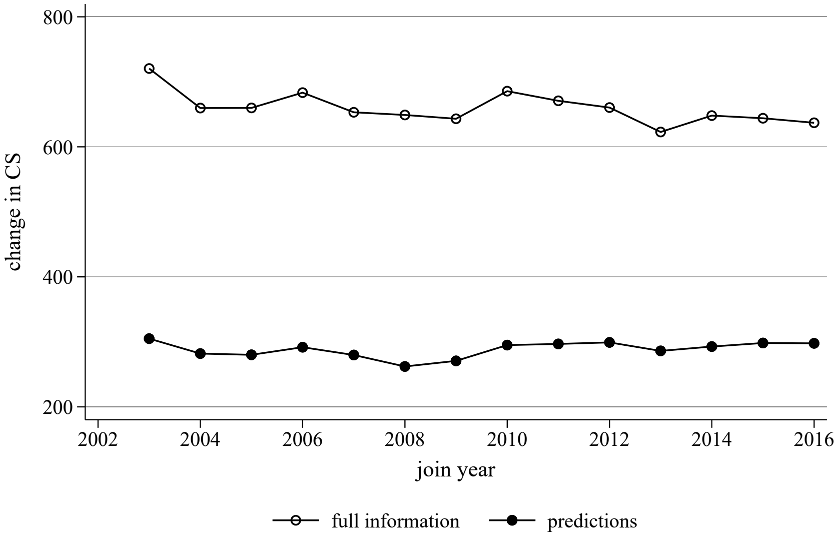


Panel A: Ownership and playtime

Panel B: Playtime at s.q. expenditure

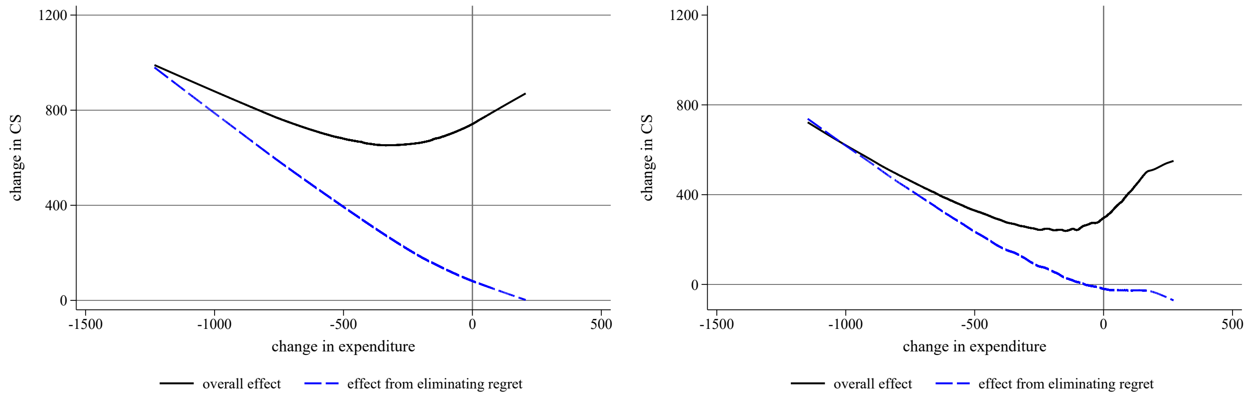
Notes: This figure shows the robustness of our descriptive results to varying utilities per hour of playtime across games. Panel A plots the share of users owning a game against the natural log of the average realized playtimes among those who own the game, for all 100 games in the sample. Panel B shows how the additional playtime that full information (or, in the dashed line, sophisticated prediction) allows status quo expenditure to achieve varies with the weight given to game-specific coefficients implied by the deviations in Panel A. Using game-specific weights that fully explain cross-game-purchase propensities, full information raises the playtime delivered by status quo expenditure by just over 60 percent, rather than the main 94 percent estimate. Full weighting reduces the effect of sophisticated prediction from 38 percent to about 25.

Figure 6: Welfare effects with varying preferences by user experience



Notes: The figure shows average changes in consumer surplus from full information (hollow circles) and sophisticated predictions (solid circles), based on logit demand parameters that are allowed to vary across user join year cohorts. That is, we calculate average ΔCS from full information and prediction, separately for each user join year.

Figure 7: Changes in consumer surplus and regret

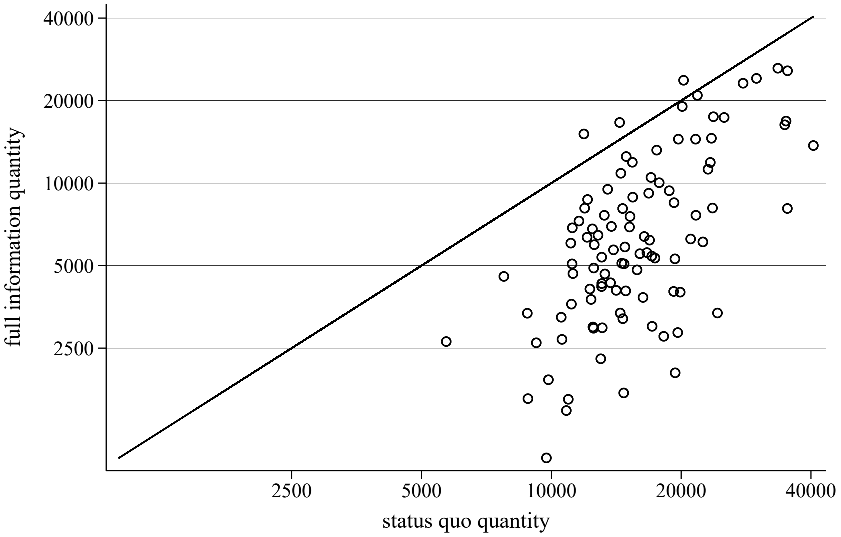


Panel A: Full information

Panel B: Predictions

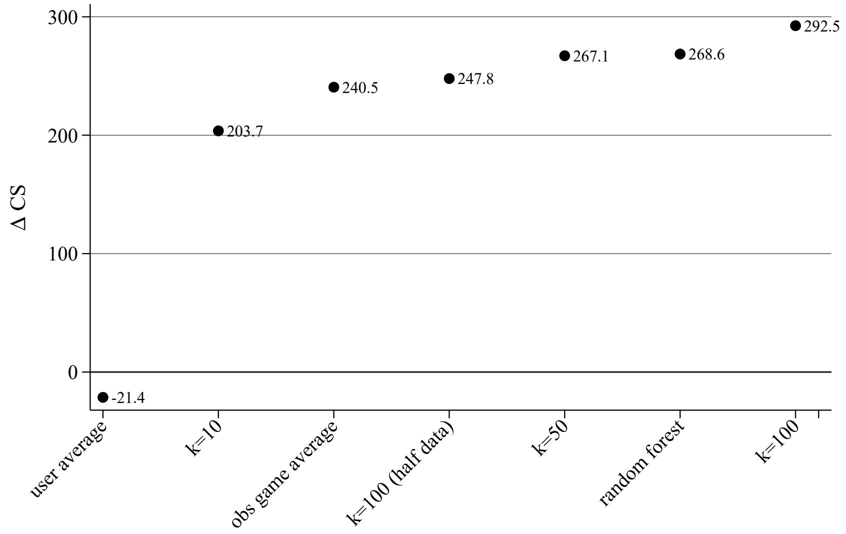
Notes: Panel A plots the smoothed relationship between individuals' changes in expenditure (on the x-axis) and their change in CS (on the y-axis) arising from full information. The blue lines show the component arising from reduced regret, while the black line shows the total change in CS. Panel B repeats the exercise for the effects of sophisticated predictions.

Figure 8: Full information vs status quo quantities by game



Notes: The figure plots status quo quantities sold (x-axis) and full-information counterfactual quantities sold (y-axis) for all 100 games in the dataset. Points below the 45-degree line indicate games selling fewer units under full information, and points above the line represent games selling more under full information.

Figure 9: Welfare effects for a progression of playtime predictions



Notes: This figure presents mean ΔCS estimates, relative to the status quo. The estimates show the increase in CS that would ensue from heeding personalized playtime predictions arising from various kinds of prediction models. The rightmost point ("$k = 100$") is the baseline estimate of the average welfare effect of our preferred prediction approach, matrix factorization with 100 latent factors.

Appendix

A Prediction approach details

In Section 4.2 we use a number of prediction approaches. Here we provide additional details about the models.

1. A common mean across users and games (an OLS regression model with just an intercept).

2. A regression model with user characteristics (these include the number of Steam "friends" a user has (ln(friends+1)), the user's average game completion rate ("average percentage of achievements earned per game"), the number of perfect games ("number of games where this player has gotten every achievement"), the time since the user joined the Steam platform, and whether the user's name is characteristically male, female, or of an unknown gender).[24]

3. A regression model with more user characteristics (the variables above plus indicators for 117 user countries of origin).

4. A regression model with a small number of game attributes: the game's price and its square, indicators for whether the game includes the indie genre tag and the action genre tag, recent and cumulative review categories (overwhelmingly positive, mostly positive, mixed, very positive, unknown), and the average review score (out of 100) for recent and cumulative reviews.

5. A regression model with a larger number game attributes. The full set of attributes consists of the above plus indicators for 77 game tags, including features such as multiplayer, strategy, sports, classic, etc.

6. Game-level average playtimes, estimated as an OLS with game fixed effects.

7. A regression model with all user and game attributes above.

8. A random forest-based prediction using all of the variables, as well as their squares and interactions (Breiman, 2001).

[24] See url's of the form `https://steamcommunity.com/id/[steamuserid]`.

9. Game and user average playtimes, estimated as an OLS with both game and user fixed effects.

10. Our collaborative filter approach, using 5, 10, 50, and 100 factors.