Quality Predictability and the Welfare Benefits from New Products: Evidence from the Digitization of Recorded Music*

Luis Aguiar†1 Joel Waldfogel‡2

[1]Institute for Prospective Technological Studies (IPTS)

[2]Carlson School of Management, Department of Economics,
University of Minnesota and NBER.

November 26, 2015

*We are grateful for comments from Avi Goldfarb and Ajay Agrawal, as well as seminar participants at the IPTS, Minnesota, the Society for Economic Research on Copyright Issues, the Summer Institute on Competitive Strategy at Berkeley, the NBER Summer Institute joint IO/Digitization meeting, the Econometric Society Meetings in Minneapolis, the 12th Conference on Media Economics in Naples, the 13th ZEW Conference on ICTs in Mannheim, the 8th ICT Conference in ParisTech, Penn State, the University of Michigan, Yale University, Stanford GSB, the University of Zurich, and Universidad Carlos III de Madrid. All remaining errors are ours. Disclaimer: The views expressed are those of the authors and may not in any circumstances be regarded as stating an official position of the European Commission, the EC Joint Research Center or the IPTS.

†E-mail: luis.aguiar@ec.europa.eu

‡E-mail: jwaldfog@umn.edu

Abstract

We explore the consequence of quality unpredictability for the welfare benefit of new products, using recent developments in recorded music as our context. Digitization has expanded consumption opportunities by giving consumers access to the "long tail" of existing products, rather than simply the popular products that a retailer might stock with limited shelf space. While this is clearly beneficial to consumers, the benefits are somewhat limited: given the substitutability among differentiated products, the incremental benefit of obscure products - even lots of them - can be small. But digitization has also reduced the cost of bringing new products to market, giving rise to a different sort of long tail, in production. If the appeal of new products is unpredictable at the time of investment, as is the case for cultural products as well as many others, then creating new products can have substantial welfare benefits. Technological change in the recorded music industry tripled the number of new products between 2000 and 2008. We quantify the effects of new music on welfare using an explicit structural model of demand and entry with potentially unpredictable product quality. Based on plausible forecasting models of expected appeal, a tripling of the choice set according to expected quality adds nearly 20 times as much consumer surplus and overall welfare as the usual long-tail benefits from a tripling of the choice set according to realized quality. We estimate that the new recorded music products raised the surplus experienced by US consumers in 2011 by $US 118 million.

Keywords: Welfare, Entry, Digitization, Recorded Music.

JEL classification: D60, L13, L82, O33.

1 Introduction

The rise of the Internet - and digitization more generally - has placed attention on the welfare benefit of cost reductions that raise the number of available products. Researchers and others have viewed the Internet as delivering infinite shelf space, allowing consumers access to a long tail of obscure products.[1] Despite the importance of long-tail effects in consumption, the welfare benefit of new products is much larger when we account for the unpredictability of product quality at the time of investment. We term this the "long tail in production."

The usual long tail idea in consumption is that the Internet allows consumers access to the large number of extant products, rather than simply the popular products that consumers might access from a local retailer with limited shelf space. While access to additional products is clearly beneficial to consumers, the benefits may be somewhat limited: given the substitutability among differentiated products, the incremental benefit of obscure products - even lots of them - can be small. A long tail in production is different. The appeal of many products to consumers is difficult to know at the time that investments are made. This unpredictability is substantial for cultural products such as books, movies, and music, leading screenwriter William Goldman to famously remark that "nobody knows anything" about which new movies will be commercially successful (Goldman, 1984). Industry observers report that roughly 10 percent of new movies are commercially successful, and the figures for books and music are similar (Caves, 2000). The unpredictability of product appeal is not limited to cultural products. Gourville (2005) reports new product failure rates between 40 and 90 percent across many categories.

When the costs of bringing new products to market fall, society can in effect take more draws from an urn of potential new products. If the appeal of new products to consumers were perfectly predictable at the time of investment, then entry of additional products would be similar to adding more shelf space, virtual or otherwise, in a retail environment. The additional products would each have limited appeal and, in particular, lower appeal than the last currently entering product. But if appeal is unpredictable - and we will confirm that it is for music - then adding more products can have substantial benefits by delivering consumers products throughout the realized quality distribution. Because product appeal is

[1] By some estimates, the benefit consumers obtain from access to a long tail of additional varieties may be as high as $1.03 billion per year for books alone in 2000 (Brynjolfsson et al., 2003).

also unpredictable in other industries, this idea may have broader applicability.

Technological change in the recorded music industry has allowed substantial growth in the number of new products in the choice set. Between 2000 and 2008 the number of new products brought to market annually tripled, leading us to ask how a tripling of the number of new products available affects welfare. We can measure the benefit as the difference between welfare with the new, enlarged choice set and a smaller choice set including a third of the recently-entering products. Yet, the welfare impact of an entry cost reduction that triples the choice set depends heavily on *which* third of existing recent products would have entered absent the cost reduction. This, in turn, depends on the predictability of quality at the time of investment. At one extreme, if product quality were perfectly predictable (the "perfect foresight" or PF case), then a reduction in the cost of entry from, say, T to T' would elicit entry of new products with expected - and realized - revenue between T and T'. The addition of these modest-appeal products to the choice set corresponds to the traditional long tail benefits. The newly entering products would necessarily raise surplus available to consumers, but the benefit might be small since none of the new products would exceed the quality of the least-attractive existing product. In the more realistic case in which quality were not entirely predictable (the "imperfect predictability" or IP case), benefits would be larger, as some new products would have high realized quality despite low expected revenue.

To quantify the benefits of new products made possible by digitization, we develop an equilibrium model of the recorded music industry that includes a structural demand model and a model of entry based on expected revenue. We use data on digital music track sales for 17 countries, 2006-2011, to estimate a nested logit model of demand. The output of the model includes both parameter estimates and measures of the realized appeal of each product, which we term δ. We use the realized δ's for the US in 2011 to develop a forecasting model of expected quality, which we incorporate in our entry model. We infer fixed costs from the expected revenue of the last entering product. The model allows us to address the two questions that motivate the paper. First, what is the effect of the cost reductions associated with digitization - which have tripled the number of products brought to market in the US - on consumer surplus and overall welfare? And second, how do these benefits, which we term the long tail in production, relate to the conventional long tail in consumption? We find that a tripling of the choice set according to expected quality adds nearly twenty times as much consumer surplus and overall welfare as a tripling of the choice set according to realized quality. That is, the long tail in production is almost twenty times as large as the traditional long tail. The new products brought about by digitization raised year-2011 consumer surplus from recorded music by $US 118 million for US consumers.

The paper proceeds in 7 sections after the introduction. Section 2 presents descriptive facts about entry in the music industry, institutions for product discovery in the digital era, and a simple model illustrating the impact of unpredictability on the welfare effects of entry. Section 3 sets out an empirical structural model of the music market. Section 4 presents the data that we will use in our estimation, while Section 5 presents our estimates of demand, expected revenue, and the fixed costs from the entry model. Next, we turn to counterfactual results in Section 6, including estimates of the main objects of interest, the welfare impact of an enlarged choice set with imperfect predictability of product appeal, both absolutely and in relation to the welfare impact of an enlarged choice set with perfect foresight. Section 7 discusses robustness of results to estimated parameters and forecasting approaches. Section 8 concludes.

2 Background

2.1 Industry Background

Since 1999, recorded music revenue has fallen by 70 percent around the world. While industry participants - particularly the major record labels - have raised concerns that declining revenue will undermine investment incentives, the number of new products brought to market has risen rather than fallen as the cost of bringing new products to market has fallen substantially. As documented elsewhere, the cost of production, promotion, and distribution of new music have fallen sharply with digitization; and the number of new recorded music products brought to market each year has risen since 1990 and more sharply since 2000 (Oberholzer-Gee and Strumpf, 2010; Handke, 2012; Waldfogel, 2013b; Aguiar et al., 2014). According to Nielsen data, the number of new music products brought to market tripled between 2000 and 2008.[2] We view this growth in the number of products as a consequence of cost reductions associated with digitization. These cost reductions are substantial enough to have enabled growth in the number of new products despite the drastic decline in revenue.

[2] See for instance http://tinyurl.com/how-many-releases.

The welfare that society derives from music depends of the benefit to consumers, beyond what they pay, plus producer surplus, less costs of production as well as product discovery. With the substantial growth in new products, we would expect product discovery costs to rise. In particular, one might expect difficulty in consumer discovery of good products among the plethora of new offerings. Indeed, it is possible that consumers would fail to discover good products among the new releases, particularly among the new products released without much fanfare (e.g. little-known artists on independent labels). Under our imperfect predictability view of the world, we would expect some of the products with low ex ante promise to be highly valuable to consumers, if they were discovered. If products with modest ex ante promise make up a growing share of the new music that becomes commercially successful, then we would infer that new products do not overwhelm the new product discovery institutions. And, indeed, in related research this is exactly what we find: products from independent labels, as well as products from new artists, make up growing and now substantial shares of the best-selling new recorded music (Waldfogel, 2013b; Aguiar et al., 2014).

What might explain these findings? In the pre-digital environment, terrestrial radio was the main means of product discovery. Music labels provided radio program directors with more music than they could air, and the program directors would choose songs to promote (sometimes with compensation). These songs were then aired to large radio audiences (Caves, 2000). The digital era has also brought some new information institutions which reduce discovery costs. The digital environment facilitates access to a great deal of information about new music, in the form of online criticism at sites like Pitchfork and aggregators such as Metacritic.[3] In addition, consumers have access to customized online "radio stations" via sources such as Pandora and Spotify. These sources reduce costs of experimentation in two ways. First, they provide informed suggestions. A consumer seeds a Pandora station with music that he or she likes; the service then presents the listener with music that resembles the seed, or is liked by people who also like the seed. Second, these suggestions are served to small numbers of individuals rather than to large audiences. One of the major social costs of product discovery is the time that listeners spend getting acquainted with new music to decide whether they like it. Playing a new song on a traditional radio station is thus a costly, large-scale experiment using the time of thousands of listeners. Serving a song via the

[3]See the discussion in Waldfogel (2013b).

Internet to targeted individuals expressing interest in related music consumes less listener time and could therefore actually be less costly.

While systematic quantification of social costs would be a useful exercise, it is not clear that changed discovery costs prevent consumers from finding good new products. Hence, we proceed with welfare analysis based directly on consumer surplus, revenue, and costs of entry, leaving aside measures of product discovery costs before and after digitization.

2.2 Related Literature

The study quantifies the benefit of a technological change that allows more entry of new products, given that new product appeal is unpredictable. Our question and approach are related to both the literature estimating the welfare effects of particular new products and the entry literature. Many studies evaluate the welfare impact of new products. A few prominent examples include Petrin (2002) and Hausman and Leonard (2002). The usual approach is to estimate demand in the presence of the new product, then to simulate welfare absent the new product. We similarly do that, but we also model the entry process. That is, the comparative static that we evaluate is not simply about whether a particular new product - such as the minivan or breakfast cereal - exists, but rather about the cost of entry that would give rise to new products.

Our paper is therefore closely related to the strand of the entry literature that incorporates demand modeling and therefore allows for explicit estimates of fixed costs (e.g. Berry and Waldfogel (1999)). Usually, researchers postulate a model in which products (or firms) enter as long as their variable profit exceeds their fixed costs; and fixed costs are estimated from the expected revenue of marginal entrants. Observed entry configurations can then be viewed as Nash equilibria given the estimated fixed costs. Such models can be used to estimate welfare under, say, counterfactual fixed costs. Our exercise does this, adding the novel feature that product appeal is unpredictable at the time of entry.

Our exercise is also closely related to the literature on the "long tail" benefits of the Internet. Brynjolfsson et al. (2003) quantify the benefit of access to the full list of books at Amazon in contrast to, say, the 100,000 books locally available to a consumer. Sinai and Waldfogel (2004) show that locally isolated consumers make greater use of the Internet. Anderson

(2006) popularized the idea of the long tail in a book asserting that the long list of products at the tail of the distribution are growing in importance relative to the small number of products at the head. All of these studies take the view - implicitly or explicitly - that digitization raises the variety available to people via an infinite shelf-space mechanism rather than the new product mechanism that we explore.

2.3 How Would Entry Cost Reduction Affect Welfare?

To fix ideas this section describes the intuition of our approach. Section 3 discusses the explicit model. When entry costs are T, then all products with expected revenue above T enter, while those with lower expected revenue do not; when the entry cost falls from T to T', then more products become viable, and more entry occurs. Having more products in the choice set raises welfare, but the size of the impact of additional products on welfare depends on the predictability of product quality at the time of investment. To see this, consider the following simple model of product entry with the possibility of quality unpredictability.

At the time of investment, a label forms an estimate of a product's marketability as the true revenue y, plus an error ν: $y' = y + \nu$.[4] If the entry cost is T, then all products with expected revenue $y' > T$ enter. If the entry cost T falls to T', then all products with $y' > T'$ enter. When product quality is perfectly predictable ($\nu = 0$), then a reduction in entry costs brings new products with expected and realized revenue - and therefore, we infer, product quality - between T and T'. In the more realistic case in which product quality is not perfectly predictable at the time of investment, the addition of products with expected revenue between T and T' elicits entry of products whose realized revenue might be anywhere in the distribution and can, of course, exceed T.

Our main concern in this paper is the evaluation of an entry cost reduction that tripled the number of new products. Given that digitization has already occurred, the welfare effect of digitization is the difference between the welfare associated with the current status quo choice set and the choice set including only a third as many new products. The major challenge to this exercise, however, is determining *which* third of recently-added status quo products would have existed if digitization had not reduced entry costs. This, in turn, depends on the

[4]This setup is reminiscent of Terviö (2009).

predictability of product quality.[5]

If investors had perfect foresight - and product quality were therefore completely predictable to investors at the time of entry - then when costs were high, only the products with the highest expected and realized quality would enter. Hence, the counterfactual high-entry-cost choice set would be the top third of products according to realized quality. The comparison of the top third of products with the total choice set is analogous to the shelf-space problem underlying the usual long tail welfare calculation asking, for example, what benefit consumers derive from access to the top million books as opposed to the top 100,000. Under this usual approach, the benefit of additional products would be relatively small. At the other extreme, if quality were completely unpredictable to investors, then the counterfactual choice set associated with high entry costs would be a random sample of status quo products. Because the additional products would be as good, on average, as existing products, the additional products would add more to welfare than if investors had perfect foresight.

In the more plausible intermediate case of imperfect predictability, the effect of new products on welfare would fall between the two polar cases. This discussion demonstrates that the impact of cost reduction on product entry and resulting welfare - the long tail in production - depends crucially on the predictability of product quality to investors.

Evaluating the welfare impact of cost reduction requires three components. First, we need a structural model of demand, which allows us to calculate the consumer surplus and revenue associated with any set of products. Second, we need a forecasting model for quality to describe the revenue that producers expect from each product at the time of investment. Third, we require an entry model that makes use of the expectations to determine the set of products entering given the entry cost structure. We can use the entry model to generate estimates of fixed costs as the expected revenue of the last entering product.

3 The Model

This section describes the components of our equilibrium model of the recorded music industry. We start by describing our structural model of demand. We then turn to our forecasting

[5] While we focus throughout the text on the welfare benefit of tripling the number of new products, we also explore the welfare consequence of different degrees of growth in the choice set. See Section 7.

model of product quality before presenting our entry model. Details of the empirical implementation are deferred until Section 5, after we introduce the data in Section 4.

3.1 Demand

Given our goal of developing an entry model incorporating expectations about product quality, we employ a model that allows us to easily infer product quality while also allowing for substitutability among products. To this end we employ a nested logit model, similar to that of Berry (1994) and Ferreira et al. (2013).

In each country, consumers choose whether to buy music and then choose among available songs. The choice sets of songs vary both across countries and over time. Define J_{ct} as the set of songs available in country c at time t, and index songs by j.[6] Suppressing the time subscript, each consumer therefore decides in each month whether to download one song in the choice set $J_c = \{1, 2, 3, \dots, J_c\}$ or to consume the outside good (not purchasing a song). Specifically, every month every consumer i in country c chooses j from the $J_c + 1$ options that maximizes the conditional indirect utility function given by:

$$\begin{aligned} u_{ij} &= x_{jc}\beta - \alpha p_{jc} + \xi_{jc} + \zeta_i + (1-\sigma)\epsilon_{ij} \\ &= \delta_{jc} + \zeta_i + (1-\sigma)\epsilon_{ij}, \end{aligned} \tag{1}$$

where δ_{jc} is therefore the mean utility of song j in country c. x_{jc} includes song as well as country-specific characteristics, p_{jc} is the price of song j in country c and α is the marginal utility of money. The parameter ξ_{jc} is the unobserved (to the econometrician) quality of song j from the perspective of country c consumers and can differ across countries for the same song (song j can for example have different quality to US vs French consumers). ϵ_{ij} is an independent taste shock. In contrast to a simple logit model, the nested logit allows for correlation in consumer's tastes for consuming digital music.[7] The parameter ζ_i therefore

[6] Our data cover only digital singles, not albums. See Section 4 for details on the data.

[7] In the logit model the individual taste ϵ_{ij} is independent across both consumers and choices and the conditional indirect utility function is given by $u_{ij} = \delta_{jc} + \epsilon_{ij}$. This prevents the possibility that consumers have heterogeneous tastes, i.e. differ in their taste for consuming music.

represents the individual-specific song taste common to all songs in the nest. Cardell (1997) shows that if ϵ_{ij} is a type I extreme value, then this implies that the error term $\zeta_i + (1-\sigma)\epsilon_{ij}$ is also a type I extreme value. The parameter σ measures the strength of substitution across songs in the choice set J_c. When $\sigma = 0$, the model resolves to the simple logit (see footnote 7) and the parameter ζ_i, the consumer-specific systematic song-taste component, plays no role in the choice decision. As σ approaches 1, the role of the independent shocks $(\epsilon_{i0}, \epsilon_{i1}, \ldots, \epsilon_{iJ})$ is reduced to zero and the within group correlation of utility approaches one. This implies that consumer tastes, while different for any consumer i across songs, are perfectly correlated within consumer i across songs.

Given the functional forms associated with nested logit, we can calculate the market share and revenue of each product for any set of product qualities δ_{jc}.[8]

3.2 Quality Prediction

The results from our demand estimation allow us to construct estimates of the mean utility of each song.[9] While our demand estimation will use data for multiple countries, we undertake the entry exercise and welfare calculations using data for only one country and year, the US in 2011, so we omit the country subscript below in discussion of the quality prediction model and the entry model. For each song j, δ_j reflects the appeal it generates for consumers based on its market share. Our model of entry with unpredictable product quality requires us to have a measure of the *expected* appeal that each song j would generate. That is, we need a measure of the appeal (or commercial success) that an investor would expect from releasing song j. For each product j in our US data, we assume that an investor contemplating the release of song j from a given artist will form a prediction of its appeal based on information available prior to release (e.g. previous sales of artist's release, time since artist's first release and the identity of the song's label):

$$\delta_j = \gamma_0 + z_j\gamma_1 + \mu_j, \tag{2}$$

[8]We also implement our estimates using a plain logit model to explore robustness of the results to the demand model specification. See Section 7 below.

[9]In the nested logit model we can calculate δ_{jc} as $\ln(S_{jc}) - \ln(S_{0c}) - \sigma \ln\left(\frac{S_{jc}}{1-S_{0c}}\right)$, where S_{jc} represents the market share of song j in country c and S_{oc} is the market share of the outside good. See Section 5 for details of implementation.

where the vector z_j contains information on song j's artist and μ_j is an error term. Note that terms included in z_j can also enter x_{jc} in the demand model (1) directly.[10] The predicted values $\delta'_j = \widehat{\gamma}_0 + z_j\widehat{\gamma}_1$ then provide us with a measure of the expected quality of each song j prior to release.

3.3 Supply and Fixed Costs

Our measure of the welfare associated with an entry configuration, or set of products that enters, is the sum of consumer surplus and revenue less the number of products times the fixed cost per product. The demand model gives us consumer surplus and revenue for any entry configuration. In order to evaluate the welfare associated with a set of entering products, we need fixed costs and the ordered set of entering products, which our entry model delivers. While the imperfect prediction model is our central approach and the approach we view as a realistic characterization, we also develop approaches using perfect foresight and no predictability, both to illustrate the intuition of our approach and to compare our estimates of the long tail in production with estimates analogous to the long tail in consumption, reflected in the perfect foresight model.

3.3.1 Perfect Foresight

Under perfect foresight (PF), products enter in order of realized quality, or δ_j. The fixed cost under the status quo is the expected revenue of the last (N^{th}) entering product.

To estimate the counterfactual perfect foresight fixed costs that give rise to one third of recent status quo entry, we must calculate the expected revenue of the last product when only the $\frac{N}{3}$ best-selling products enter. To this end, define δ_j as the realized quality of product j, and define Δ_j as the set of products $\{\delta_1, \ldots, \delta_j\}$. Because products are imperfect substitutes, revenue to each product depends on the full set of products in the market. The expected revenue to product 1 entering alone depends on Δ_1, and so on. That is, if $\mathrm{E}[r_k]$ is the expected revenue of product k, then $\mathrm{E}[r_k]$ is a function of the vector Δ_k.

If we order the products such that $\delta_k > \delta_{k+1}$, the products enter as long as $\mathrm{E}[r_k(\Delta_k)] > T$.

[10] We note here that some important variables in z_j, namely the information on record labels, are available only for the US, necessitating separate estimation of demand and prediction models. See Section 5.

For example, given the nested logit structure, the expected and realized revenue to product 1 when it is alone is

$$r_1 = pMs_1 = pM\left[\frac{e^{\frac{\delta_1}{1-\sigma}}}{D_1^\sigma + D_1}\right], \tag{3}$$

where $D_1 = e^{\frac{\delta_1}{1-\sigma}}$, p is the price of the product, and M is market size.[11]

More generally the revenue to product k (when it is the last entering product) is given by

$$r_k = pMs_k = pM\left[\frac{e^{\frac{\delta_k}{1-\sigma}}}{D_k^\sigma + D_k}\right], \tag{4}$$

where $D_k = \sum_{j=1}^{k} e^{\frac{\delta_j}{1-\sigma}}$. To estimate counterfactual fixed costs when $\frac{N}{3}$ products enter, we can infer that the fixed costs (T) equal the expected (and realized) revenue of the last entering product: $T \approx r_k, \quad k = \frac{N}{3}$.

Our PF fixed cost estimates require an important caveat (which applies to our imperfect predictability estimates as well). We derive our estimates of fixed costs from the expected revenue of the marginal entering product. Hence, strictly speaking, our fixed cost is an estimate of the fixed cost for the marginal entrant. It seems likely that infra-marginal entrants incur higher fixed cost. This means, further, that our estimate of the aggregated fixed costs incurred by all entrants, $N \cdot FC$, is a lower-bound on the resources consumed by the fixed costs of entry. Underestimation of $N \cdot FC$ would lead to over-estimates of welfare. We can, however, place an upper bound on fixed costs as well. Under free entry, entry could occur until profit opportunities have been dissipated. Hence, total revenue itself provides an upper-bound estimate of aggregate fixed costs ($N \cdot FC$). See Section 5.1.3.

3.3.2 No Predictability

At the opposite extreme from the perfect predictability model is a model with no predictability. While not a plausible depiction of reality, this model nevertheless provides a useful benchmark, describing a world in which, literally, "nobody knows anything." With

[11] In our empirical implementation, we define the market size as 12 times the country population.

no predictability, all products are identical ex-ante. Hence the expected revenue of any product depends only on the total number of products entering (k) and is the total revenue to those k products divided by k. That is,

$$\mathrm{E}\left[r_k\right] = pM\mathrm{E}\left[\frac{\frac{D_k}{D_k^{\sigma}+D_k}}{k}\right], \tag{5}$$

where D_k is evaluated with a particular draw of k product qualities (δ_j), p is the price and M is market size.

Hence, the no prediction estimate of status quo fixed cost is the total observed revenue divided by the number of products. We estimate counterfactual fixed cost as the average revenue per product if $\frac{N}{3}$ products entered. To estimate this, we take draws of $\frac{N}{3}$ δ's, and each draw generates an estimate of average revenue per product.

Under the no predictability model, additional products add substantially to welfare by construction because the average quality of products does not decline with entry. The only reason that consumer surplus and the expected revenue per product decline with entry is through substitution allowed for by the nested logit model's parameter σ.

3.3.3 Imperfect Prediction

The perfect foresight and no-prediction models present two extremes, both somewhat unrealistic. This leads us to the imperfect prediction case, in which investors have some ability to predict the appeal of songs at the time of investment. Our predicted δ's (which we term δ') create an ordering of potential projects in descending order of ex ante (expected) promise: $\delta'_1 > \delta'_2 > \ldots > \delta'_N$. In the no prediction case (above), we took a random draw of the k products to estimate the revenue per product when k products enter. In the imperfect prediction case, the analog to a random draw of k products is the top k products ordered by expected quality.

We calculate the expected revenue of the k^{th} entrant as follows. Order songs by their ex ante promise (δ'). When the first $(k-1)$ songs, ordered by their ex ante appeal, are in the market with their ex post appeal, the revenue to the k^{th} entrant depends on its realized value. For a particular realization of $\delta_k = \delta'_k + \varepsilon$, the share of population consuming product k, via the nested logit formula, is:

$$s_k(\varepsilon) = \frac{e^{\frac{(\delta'_k+\varepsilon)}{1-\sigma}}}{\left[\sum_{j=1}^{k-1} e^{\frac{\delta_j}{1-\sigma}} + e^{\frac{(\delta'_k+\varepsilon)}{1-\sigma}}\right]^{\sigma} + \left[\sum_{j=1}^{k-1} e^{\frac{\delta_j}{1-\sigma}} + e^{\frac{(\delta'_k+\varepsilon)}{1-\sigma}}\right]}. \tag{6}$$

Because of the nonlinearity of the share formula, we compute the expected market share by integration. The expected market share of the k^{th} entrant is therefore given by

$$\mathrm{E}\left[s_k\right] = \int s_k(\varepsilon) f(\varepsilon) d\varepsilon, \tag{7}$$

where f is the density of ε. In our empirical implementation, we will take f to be the empirical distribution of the residuals from our prediction model, $\varepsilon \equiv \delta - \delta'$. We will therefore compute the expected revenue of the k^{th} entrant (when the first $(k-1)$ songs ordered by their ex ante appeal have entered) as

$$\mathrm{E}\left[r_k\right] = pM\mathrm{E}\left[s_k\right] = pM\left[\frac{1}{N}\sum_{n=1}^{N} s_k(\varepsilon_n)\right] = pM\left[\frac{1}{N}\sum_{n=1}^{N} \frac{e^{\frac{\delta'_k+\varepsilon_n}{1-\sigma}}}{D_{kn}^{\sigma} + D_{kn}}\right], \tag{8}$$

where $D_{kn} = \sum_{j=1}^{k-1} e^{\frac{\delta_j}{1-\sigma}} + e^{\frac{\delta'_k+\varepsilon_n}{1-\sigma}}$ and N is the total number of products.

We estimate status quo fixed costs using the expected revenue of the last entrant, and we estimate counterfactual fixed cost as the expected revenue of the last $(\frac{N}{3}^{rd})$ product when the top $\frac{N}{3}$ products enter according to expected quality, or $k = \frac{N}{3}$.

4 Data

Given our goals of estimating the welfare benefits of new music products, we would ideally observe all revenue generated by new music products. This would include sales of digital music, sales of physical products (e.g. CD's) as well as live performance revenue. Our actual data, while very rich and detailed, include only a subset of the ideal. That is, the basic data for this study include annual sales of all digital singles in the US, Canada, and 15 European countries, 2006-2011, but our data contain no information on physical products nor live performance revenue.[12] Our sample includes 3,984,227 distinct tracks from 75,235 distinct artists and, because a song can appear in multiple countries and years, 50,828,216 observations. Total digital track sales in the data are 628.3 million in 2006 and rise to 1512.4 million in 2011.

The sales data are drawn from Nielsen's SoundScan product, which serves as "a major source for the Billboard charts and is widely cited by numerous publications and broadcasters as the standard for music industry measurement." Nielsen tracks what consumers are buying "both in-store and digitally." In particular, they "compile data from more than 39,000 retail outlets globally."[13] We use the same version of the Nielsen data employed in Aguiar et al. (2014), and readers are directed there for details on the dataset construction.[14]

We use these underlying data to create two datasets that we use for demand estimations and quality predictions, respectively. The demand estimation dataset covers 17 countries 2006-2011 and includes data on artists' country of origin, age (time since first release) as well as an artist genre designation. We obtained the genre data from Allmusic.com.[15] We perform the quality prediction exercise and welfare calculations using only the subset of US data. These data also include the identity of labels releasing each song. Our revenue data cover digital track sales, not the total revenue that artists earn from creating music. The track sales understate total recorded music sales. Our US digital track sales total \$1.313 billion for 2011, while the RIAA reports total recorded music sales of \$7.008 billion.[16] Hence, to make them reflective of US recorded music sales, we scale up our estimates by 5.34.[17] We discuss this further in Section 5.3 below.

[12]The dataset initially includes the following 16 European countries: Austria, Belgium, Denmark, Finland, France, Germany, Ireland, Italy, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland and the United Kingdom. However, given that Poland enters the data in 2008 only, we decided to drop it from the analysis.

[13]http://www.nielsen.com/content/corporate/us/en/solutions/measurement/music-sales-measurement.html.

[14]That dataset excludes entries that appear not to be songs and includes only artists whose national origins can be determined from MusicBrainz (www.musicbrainz.org). The latter criterion excludes 44.4 percent of otherwise valid observations while retaining 91 percent of sales.

[15]We sought matches for each of the 75,235 sample artists from Allmusic.com. We obtained matches for 61,073 artists, accounting for 93.2 percent of the sales in the data with origin matches. The artists are clasified into 36 distinct genres.

[16]The RIAA reports sales of 1,306.2 million digital tracks, generating \$1,492.7 in revenue, or \$1.14 per track. Our data contain 1.149 billion US track sales. At \$1.14 per track, our data cover \$1.313 billion in track sales. See RIAA, 2011 Year-End Shipment Statistics.

[17]Artists also derive revenue from live performance as well as recorded music. In 2011, live performance revenue was \$4.35 billion. See 2011 Pollstar Year End Business Analysis, available at http://www.pollstarpro.com/files/Charts2011/2011BusinessAnalysis.pdf. To the likely extent that the creation of new music also brings the opportunity to generate some live performance revenue, measures of expected revenue based only on recorded music sales would understate the true expected revenue.

We also employ measures of the digital share of music expenditure in each country and year. These are obtained from the recording industry's international umbrella trade organization, the International Federation of the Phonographic Industry (IFPI).

5 Empirical Implementation

5.1 Demand Model

We now turn to the empirical implementation of our equilibrium model of the recorded music industry. We start by presenting the estimation of our structural demand model. We then present, in turn, our forecasting model of product quality and the estimation of the fixed costs of entry.

Following equation (1) and normalizing the utility of the outside good δ_{0c} to 0, the market shares for all $j \in J_c$ are given by $S_{jc} = \frac{e^{\frac{\delta_{jc}}{1-\sigma}}}{D_{J_c}^{\sigma} + D_{J_c}}$, where $D_{J_c} = \sum_{j \in J_c} e^{\frac{\delta_{jc}}{1-\sigma}}$. Inverting out δ_{jc} from observed market shares as in Berry (1994) yields

$$
\begin{aligned}
\ln(S_{jc}) - \ln(S_{0c}) &= \delta_{jc} + \sigma \ln\left(\frac{S_{jc}}{1 - S_{0c}}\right) \\
&= x_{jc}\beta - \alpha p_{jc} + \sigma \ln\left(\frac{S_{jc}}{1 - S_{0c}}\right) + \xi_{jc}, \qquad (9)
\end{aligned}
$$

so that an estimate of β, α and σ can be obtained from a linear regression of differences in log market shares on product characteristics, prices and the log of within group share. The estimate of σ will be positive if variation in a song's share relative to the total inside share $(1 - S_{0c})$ explains $\ln(S_{jc}) - \ln(S_{0c})$ conditional on the other explanatory variables. Here, x_{jc} includes a constant, the artist's national origin, the age of each song, its genre, a set of year dummy variables, and a host of country level controls. In particular, we include a direct measure of digital share of music expenditure in each destination and year, GDP per capita, the urban share of the total population, the percentage of fixed broadband Internet subscribers, the percentage of mobile cellular subscriptions, and the percentage of Internet users.

5.1.1 Identification of σ

Because the inside share of each song j is, by construction, endogenous in equation (9), we need to use instruments in order to consistently estimate σ. For this, we follow the common practice in demand estimation of instrumenting for the inside share using functions of other products' characteristics (Berry et al., 1995; Verboven, 1996). In our application, these include the number of songs by country, the sums of the number of songs by origin and genre, as well as the sums of other products' ages. The simplest version of this type of instrument is the sum of the number of products by country and year.

5.1.2 Price coefficient

Ideally, we would observe exogenous price variation across songs that would allow us to econometrically identify the price coefficient α. This approach is unfeasible because we do not observe song-level prices. We do, however, observe the average price, allowing to infer the α parameter from a first-order condition on pricing.

Because the price is constant, the term αp_{jc} in (9) simply becomes part of the constant term in estimating equation

$$\ln(S_{jc}) - \ln(S_0) = x_{jc}\beta + \sigma \ln\left(\frac{S_{jc}}{1 - S_{0c}}\right) + \xi_{jc}. \tag{10}$$

Using σ we can calculate the country-specific mean utility of each song δ_{jc}:

$$\delta_{jc} = \ln(S_{jc}) - \ln(S_{0c}) - \sigma \ln\left(\frac{S_{jc}}{1 - S_{0c}}\right). \tag{11}$$

We can infer α from a condition on the music demand elasticity. Assuming that songs are sold by a profit maximizing monopolist facing zero marginal cost, the price level would be set such that the demand for songs is unit elastic.[18] Given that the elasticity of demand for music in our model is given by $\eta = \alpha p \left[1 - \frac{D_J}{D_J + D_J^\sigma}\right]$, we can infer the price parameter under

[18]Note that this way of inferring α is not uncommon among practitioners. As noted by Björnerstedt and Verboven (2013), one may want to verify whether elasticities are consistent with external industry information as opposed to relying too heavily on econometric estimates. While our motivation is driven by lack of data on product prices, we basically follow the same type of approach.

the assumption of unit-elastic pricing as $\alpha = \frac{1}{p}\frac{D_J + D_J^\sigma}{D_J^\sigma}$.

In reality, it is likely that major sellers of digital music (e.g. Apple) price songs below the static profit maximization level to stimulate demand for complementary hardware (Shiller and Waldfogel, 2011; Danaher et al., 2014). If so, the estimate of α is an upper bound, and our resulting estimates of consumer surplus will be a lower bound.[19]

At this point we therefore have estimates of σ, α and mean utilities (δ_j) for each product, which allow us to calculate consumer surplus and revenue.

5.1.3 Consumer Surplus, Revenue, and Welfare Measures

Given our estimates of σ and α, we can calculate the mean utility of each song, and given these estimates of δ_{jc} we can calculate the consumer surplus (CS) and revenue (Rev). These, in turn, allow us to calculate two kinds of welfare measures, CS and overall welfare $W = CS + Rev - N \cdot FC$, where N is the number of products and FC is the fixed cost per product. Note that if entry costs equaled revenue, then welfare would simply equal consumer surplus. In what follows, we calculate the change in welfare both assuming that fixed costs are determined by the marginal entrant as well as under the assumption that fixed costs equal revenue, in which case $\Delta W = \Delta CS$. Use of consumer surplus as a welfare measure is also consistent with the literature in this area (e.g. Brynjolfsson et al. (2003)) which focuses entirely on CS. Consumer surplus is given by[20]

$$CS = \frac{M}{\alpha} \ln \left(\sum_J D_J^{1-\sigma} \right) = \frac{M}{\alpha} \ln \left(D_J^{1-\sigma} + 1 \right). \tag{12}$$

Revenue is given by

$$Rev = p_j M \left[\frac{D_J}{(D_J^\sigma + D_J)} \right]. \tag{13}$$

Our two main objects of interest are the absolute change in welfare with the new products

[19]If demand is inelastic, then $p\alpha \left[1 - \frac{D_J}{D_J + D_J^\sigma}\right] < 1$ and $\alpha < \frac{1}{p}\frac{D_J + D_J^\sigma}{D_J^\sigma}$.

[20]The results from our estimations allow us to calculate CS and revenue for each country in each year. However we omit the country and time subscripts since we perform our counterfactual exercise on US data in 2011 only.

and the change in welfare under our IP approach, relative to the standard PF long tail. Given our setup, $\Delta CS_{IP} = \frac{M}{\alpha}\left[\ln(D_J^{1-\sigma}+1) - \ln(D_{J_0^{IP}}^{1-\sigma}+1)\right]$, where J is the full status quo choice set, J_0^{IP} is the set of products that would have existed absent cost reduction under IP, and $D_J = \sum_{j=1}^{J} e^{\frac{\delta_j}{1-\sigma}}$. Note that this depends on α, σ, and our predictions of which products enter the counterfactual IP choice set. $\Delta W_{IP} = \Delta CS_{IP} + \Delta Rev_{IP} - N_0 \cdot FC_0 - N_{IP} \cdot FC_{IP}$, where ΔRev_{IP} is the status quo revenue less the revenue that the top third of products in expected revenue would generate, and N_0 and FC_0 are the number of products and the fixed costs per product in the status quo, respectively.

Our second main objects of interest are the welfare change ratios:

$$\frac{\Delta CS_{IP}}{\Delta CS_{PF}} = \frac{\frac{M}{\alpha}\left[\ln(D_J^{1-\sigma}+1) - \ln(D_{J_0^{IP}}^{1-\sigma}+1)\right]}{\frac{M}{\alpha}\left[\ln(D_J^{1-\sigma}+1) - \ln(D_{J_0^{PF}}^{1-\sigma}+1)\right]} = \frac{\ln\left(\frac{D_J^{1-\sigma}+1}{D_{J_0^{IP}}^{1-\sigma}+1}\right)}{\ln\left(\frac{D_J^{1-\sigma}+1}{D_{J_0^{PF}}^{1-\sigma}+1}\right)}, \tag{14}$$

and, analogously, $\frac{\Delta W_{IP}}{\Delta W_{PF}}$. These ratios depend on σ and the products predicted to enter the choice set. Notice that while $\frac{\Delta W_{IP}}{\Delta W_{PF}}$ depends on α, the ratio $\frac{\Delta CS_{IP}}{\Delta CS_{PF}}$ does not.

5.1.4 Results

Table 1 reports estimates of the demand models. In order to keep the data to a manageable size, we rely on a randomly drawn 5% subset of our full data set. The results presented below therefore rely on a sample containing 2,384,157 observations. Columns (1)-(5) correspond to one-level nested logit models estimated on data for 17 countries, 2006-2011, including song age and year dummies as controls. Column (1) presents OLS results and an estimate of σ close to 1, reflecting the fact that we are regressing a function of $\ln(S_{jc})$ on another function of $\ln(S_{jc})$. Specifications (2)-(5) use the sum of age, genre, and origins by markets (country and year) as instruments to consistently estimate σ, and add characteristics used as controls. The coefficient of greatest interest in these estimates, σ, varies between 0.607 in column (2) and 0.519 in column (5). The first stage F-statistics for excluded instruments in column (2)-(5) vary between 250.28 and 829.48, respectively.

We also estimated a two-level nested logit model using genres as nests. See column (6). We do not reject the hypothesis that the parameter measuring substitutability across genres equals the parameter measuring substitutability within genre, so we proceed with the one-level nested logit model throughout. In particular, we employ the rich specification in column (5) as our baseline model, but we also explore sensitivity of results to different values of σ in Section 7.3.

5.2 Quality Prediction

While we estimate the demand model on data for 17 countries over the period 2006-2011, we perform our counterfactual exercises on only US data for 2011. In our counterfactual calculations we treat the vintage-2011 products as endogenous. That is, we treat the pre-2011 products available in 2011 as exogenously available and omit the bottom two thirds of vintage 2011 products (according to their expected quality) in our counterfactual choice set. This simulation can be interpreted to represent a cost reduction that occurred starting in 2011.

To predict the quality of new products in 2011, we can use the information available on the artist associated with song j. Define z_j as the vector containing information on song j's artist (e.g. artist's past sales, etc.). Because our counterfactual exercise is performed on US data for the year 2011, we can regress δ_j for the 2011 releases on their characteristics at release, that is, the z_j's for the 2011 songs and then to use the fitted values of δ_j as our measure of expected quality. This approach faces a complication: a regression of the vintage-2011 δ_j's - which contain the realized qualities of songs in 2011 - uses information not available prior to the release of the songs. That is, if $\delta'_j = z_{j,2011}\widehat{\gamma}_{2011}$, the $\hat{\gamma}$ from the 2011 regression contains the realizations for vintage-2011 δ_j, which were not known when the 2011 songs were released. This challenge can be overcome by using data available prior to release. For example, if we estimate $\widehat{\gamma}$ from a regression of quality realizations for vintage-2010 δ_j on $z_{j,2010}$ (i.e. characteristics in year 2010), the resulting $\widehat{\gamma}_{2010}$ can be applied to characteristics of vintage-2011 songs to produce a prediction of the vintage-2011 songs' qualities in 2011: $\delta'_j = z_{j,2011}\widehat{\gamma}_{2010}$ that only uses information available prior to release.

Ideally, the quality forecasting model would include all variables predictive of success that are known to investors prior to release. While there is much that we do not observe - such as the characteristics of the artist's music and appearance - we do observe some important information. We observe artists' genres and countries of origin. For artists who are not new, we observe past sales of their previous songs, which may be predictive of the sales of new work. We also observe the artist's "age" (time since first release vintage) as well as the time since the last release by the artist (prior to the current release). Importantly, we also observe the identity of the label releasing the song. The data contain 13,507 different labels. Artists tend to match with different labels according to expected quality, with the "major" labels releasing artists with substantial commercial appeal and the independents releasing artists with more modest prospects. There is, moreover, a range of independent labels from labels such as Merge and 4AD handling well-known "indie" artists to more obscure labels. Hence, label dummies should be correlated with predictors of success that labels can observe but the econometrician cannot.[21]

Our first task is to show that we have pre-release variables that are predictive of the success of a release (i.e. predictive of δ). To this end, Table 2, column (1) reports a regression of δ_j for vintage-2010 releases in 2010 on the songs' artists' past sales, in years 2006-2009, along with terms in artist age, an indicator variable for new artists, and time since last release.[22] While our goal here is forecasting rather than inference about particular parameters, it is comforting that coefficients have intuitive signs. Artists with greater past sales have higher δ's. Older artists - those whose first release was longer ago - have lower sales. Artists whose last release occurred earlier have higher sales, perhaps reflecting more pent-up demand. The R^2 of the expression is 0.196, and the R^2 of the prediction (the square of the correlation of realized δ's with their predictions) is 0.200. That explanatory variables matter - and that R^2 exceeds zero - means that it is not strictly correct that "nobody knows anything" about which products will succeed. The second column adds origin and genre fixed effects, raising regression and prediction R^2 to 0.225 and 0.213, respectively. The third column adds label fixed effects for all labels with more than 250 US releases in 2011, raising regression R^2 to 0.403 and prediction R^2 to 0.325. The rather large deviation between the regression and the prediction R^2 in column (3) suggests that model (3) overfits the data and that the regression in column (3) overstates the model's actual predictive ability. Column (4) reports the column (3) regression on 2011 data rather than 2010. The fitted values of this regression provide "predictions" of 2011 product quality that are more accurate, in sample, than the forecasts generated from the 2010 regression. The column (4) regression R^2 of 0.411 far exceeds the

[21] To the extent that labels have already formed a prediction of the artist's appeal when signing them, including label fixed effects in the forecasting model will arguably lead to more conservative results.

[22] We use 2010 quality realizations rather than 2011 realizations because we use this for the 2011 quality prediction.

prediction R^2 measures derived from 2010 models. Although we believe the column (4) model overstates predictive ability, greater accuracy in prediction leads to conservative estimates of the welfare benefit of new products. We therefore use these predictions for examining robustness of our results to better prediction ability in Section 7.4.

We make one final observation. Our prediction model tells which vintage 2011 products would not have been available to consumers in 2011 absent cost reductions following from digitization. In reality, cost reduction - and the growth of new releases - predate 2011, so we will need to bear this in mind in order to develop estimates of the benefit of all new products since digitization that consumers experience from consumption during 2011.

5.3 Fixed Costs

Our estimates of fixed costs are based on estimates of the expected revenue of the last entering US product. That is, we calculate perfect foresight status quo fixed costs as the expected (and realized) revenue of the lowest-appeal vintage 2011 product. Because the lowest revenue observed in the US digital song data for a vintage 2011 song is \$1.14, the resulting status quo fixed cost estimate under perfect foresight is \$1.14. Scaling this up to the total year-2011 US recorded music revenue (multiplying by 5.34) yields a fixed cost of \$6.09 (see Table 3). The analogous perfect foresight counterfactual fixed cost is estimated as the expected revenue of the last entering vintage-2011 product when all pre-2011 products are in the choice set while only the top third of vintage-2011 products (by realized quality) enter. We estimate this to be \$133.96.

We estimate status quo no predictability fixed costs under the 2011 approach by calculating the average revenue to each of the vintage-2011 products when they are available alongside the earlier, exogenous products (from vintages prior to 2011). We estimate these as \$9,468. For the counterfactual no predictability fixed costs, we randomly remove two thirds of the vintage 2011 songs, then calculate the average revenue per 2011 song when, again, they are sold alongside all of the pre-2011 songs. We repeat this random exercise 5,000 times, resulting in a counterfactual fixed costs estimate of \$10,336.

We calculate the imperfect predictability status quo fixed costs by ordering the 2011 products by expected quality. We then seek an estimate of the expected revenue of the last entering vintage-2011 product when it is available alongside both the preceding vintage-2011 product and all of the pre-2011 products. Using equation (8), we estimate the status quo fixed cost as the expected revenue of the last entering product. We obtain an estimate of \$26.57. We similarly estimate the counterfactual imperfect predictability fixed costs as the expected revenue of the $k = (\frac{N}{3})^{rd}$ entering product, obtaining an estimate of \$2,516 when scaled up.

As is customary in the empirical entry literature, our fixed cost estimates are derived from a cross section of revenue data. The fixed costs derived from year-2011 expected revenue of new vintage-2011 songs reflect only expected first-year song revenue. If first year revenue is proportional to lifetime revenue, then our fixed cost estimates will be proportional to the true underlying fixed costs. Moreover, the fixed cost estimates derived from first-year revenue bear the same relationship to total fixed costs that our observed first-year revenue bears to total revenue. Hence, revenue and cost estimates are consistent with one another, for example for the purpose of entry counterfactuals involving different fixed cost levels.

For some purposes, however, one might want to adjust our fixed cost estimates. For instance, one might want to compare our fixed cost estimates to outside estimates of the cost of bringing new music to market. Artists and labels release products in order to earn *all* of the revenue that those products can generate. In addition to first-year recorded music revenue, there is the additional revenue from the remaining life of the song. Analysis of sales by time and vintage shows that the revenue generated in the first year of a song's life accounts for an average of 18 percent of lifetime song revenue.[23] Hence, we could further inflate first-year revenue by an additional factor of 5.46 (1/18 percent) to yield estimated fixed costs from the status quo IP model of the expected lifetime total recorded music revenue. This would, correspondingly, give rise to a larger estimate of the fixed cost of entry. Beyond the lifetime recorded music revenue is also live performance revenue, which reached \$4.35 billion in the US in 2011.[24] If based on first year recorded music revenue, and if live performance revenue follows similar patterns, the first-year revenue from live performance and recorded music together is $26.57 \cdot (1 + \frac{4.35}{7}) = \$43.08$. When these first year revenue sources are scaled to lifetime revenue, this becomes $\$43.08 \cdot 5.46 = \$235.22$. In what follows we inflate to total recorded music revenue, but we note that all revenue sources, together, are relevant if one

[23] A regression of the log of s_{tv} (the share of year-t sales originally released at vintage v) on age dummies and vintage dummies allows us to infer the share of sales by age from the coefficients on age dummies. Using this approach, as in Waldfogel (2012), we find that 18 percent of lifetime sales occur during the calendar release year. See Appendix B for details.

[24] See Pollstar document cited in footnote 17.

wanted to assess the realism of our implied fixed costs estimates.

While the status quo and counterfactual fixed costs estimates are mainly inputs into our welfare calculations, they are also of some direct interest as answers to the question "how much must fixed cost have fallen to generate a tripling of entry?" The answer, under imperfect predictability, is a factor of one hundred, from $2,516 to $27 in Table 3, or two orders of magnitude.

6 Simulations

We now turn to evaluating the welfare benefits of tripling the choice set.

6.1 Effect of Tripling the Number of Songs on Welfare

Table 4 reports baseline estimates of our main objects of interest, the absolute changes in welfare measures (ΔCS and ΔW) as well as the ratios $\frac{\Delta CS_{IP}}{\Delta CS_{PF}}$ and $\frac{\Delta W_{IP}}{\Delta W_{PF}}$. Using our imperfect predictability approach, the additional vintage-2011 songs in the 2011 choice set raise CS by $27 million in 2011, and given the implied reduction in entry costs, W rises by $166 million. Recall that these are inflated to reflect total US recorded music sales.

The perfect foresight welfare benefits of additional entry, corresponding to the traditional long tail, are far smaller than the IP benefits. CS for 2011 rises by $1.38 million with a tripling in the number of new 2011 products, while W rises by $8.03 million. As a result, our long tail in production produces a ΔCS_{IP} benefit that is 19.84 times larger than traditional perfect foresight benefit ΔCS_{PF}. Our overall welfare benefit ΔW_{IP} is 20.63 times larger.

Consumers experienced a $27 million benefit in 2011 from the vintage-2011 products made possible by digitization. During 2011, US consumers also enjoyed additional new products released in 2010, 2009, 2008, and so on, back to 2000 if one were to mark the onset of digitization following Napster. It would be useful to have an estimate of the additional benefit that consumers experience in 2011 from all of the songs in the 2011 choice set made available by digitization. Define n_v as the number of new (digitally enabled) songs from vintage v, and define s_v as the share of year-2011 sales of all songs from vintage v. We know that the new digitally enabled vintage-2011 songs account for $27 million in consumer surplus. If the digitally enabled songs of previous vintages v are on average as valuable as the vintage-2011 songs at release, then the contribution of vintage-v songs to year-2011 CS should be roughly proportional to the vintage-2011 contribution. Then we can estimate the contributions of earlier vintages to year-2011 consumer surplus as $\Delta CS_v = \frac{n_v}{n_{2011}} \frac{s_v}{s_{2011}} \Delta CS_{2011}$. If the vintages since 2000 include the digitally enabled new songs, then we can estimate the total benefit of these songs by inflating the original $27 million by: $\sum_{v=2000}^{2011} \frac{n_v s_v}{n_{2011} s_{2011}}$.

We can observe N_v, the total number of new products from each vintage, directly from the 2006-2011 sales data (total including both digitally enabled and others). We can infer it for earlier years from the number of older products selling during 2006-2011.[25] We can then estimate n_v for each vintage as the number released in each year less the number released in the last pre-digitization year. That is, $n_v = N_v - N_{1999}$. We estimate the inflation factor to be 4.31. This is a rough estimate for a variety of reasons, including that consumer surplus is not linear in the number of products, due to decreasing marginal utility. Still, the full year-2011 benefit of new products is roughly four times the benefit arising from just the new (vintage-2011) products. This is $118 million for the US in 2011, a year in which total recorde music sales were $7 billion.

It is useful to compare our estimates of the benefits from new products to the existing long tail literature. Brynjolfsson et al. (2003) estimate that access to all book titles at Amazon, rather than just the top 100,000 titles, delivered $1 billion in additional consumer surplus to US consumers in 2000. Their measurement approach corresponds to what we term perfect foresight but applied to all vintages rather than just the 2011 vintage. Our basic PF approach counterfactually removes the lowest-demand two thirds of products released in 2011. We can produce an estimate more closely resembling Brynjolfsson et al. (2003)'s approach by discarding all but the 100,000 most popular products among the full 2.2 million products available in 2011 regardless of vintage. The loss in CS from eliminating all but the top 100,000 products is $166 million. This figure remains smaller than the corresponding measure for books, largely because books have far lower sales concentration. Brynjolfsson et al. (2003) report that books outside the top 100,000 titles accounted for about 40 percent of book sales

[25]In particular, we estimate N_v by regressing $\ln N_{tv}$ the log of the number of products from vintage v selling year t on dummies for age (year less vintage) and dummies for vintage. Exponentiated vintage dummies then provide our estimates of N_v. These estimates stand at 30,000 in 1999, hover around 45,000 until 2002, then rise: to 78,000 in 2003, to 156,000 in 2005, and reach a peak of 157,000 in 2009.

in 2000. In our music data, tracks outside the top 100,000 account for under 5 percent of sales. Hence, we expect our estimates of conventional long tail benefits (the benefits arising from access to products outside, say, the top 100,000) to be much smaller than a corresponding estimate for books.

7 Robustness

Our estimates of the absolute changes in welfare as well as the ratios such as $\frac{\Delta CS_{IP}}{\Delta CS_{PF}}$ depend on a host of underlying model features, including the substitutability of products in the demand model (σ), the ability of investors to forecast quality at the time of investment, and the magnitude of the enlargement of the choice set (the share of status quo products available in the higher-cost counterfactual - one third in the default). In this section we consider the sensitivity of our estimate to these modeling decisions. While the ΔCS ratio does not depend on the price parameter α, the absolute welfare changes do, as does the ΔW ratio.

7.1 The Price Parameter

While the price parameter α has no effect on the ΔCS ratio, it has a direct effect on the absolute measure ΔCS_{IP}. Here, we consider the ΔCS_{IP} estimates resulting from a range of α estimates. Our baseline α is a bound derived from assuming revenue maximizing song pricing. If we instead assumed that prices were set such that the elasticity of demand were one half rather than one, then α would be half as large, and ΔCS_{IP} would be double from its baseline of \$27.35 million to \$54.70 million, meaning that \$27.35 million is a lower-bound estimate of the change in surplus from the new vintage-2011 songs in 2011. By contrast, if prices were set such that the elasticity were two, then ΔCS_{IP} would be half its baseline value, or \$13.68 million.

7.2 Share of Products Included in the Counterfactual Choice Set

Our baseline counterfactual is a world in which all old - and only one third of vintage-2011 - products exist. It is useful to know how the ratio of interest varies for different counterfactual shares that correspond to different amounts of growth in the choice set besides tripling. To this end, we re-estimate ΔCS_{IP} and $\frac{\Delta CS_{IP}}{\Delta CS_{PF}}$, including different numbers of vintage-2011 products in the counterfactual scenarios. Our baseline tripling of products corresponds to 0.33 along the horizontal axis in Figure 1. The more products that entry cost reduction is assumed to effect (i.e. the lower the share of products included in the counterfactual choice set), the larger the growth in absolute welfare measures in Figure 1. For example, if the number of products introduced in 2011 increased by a factor of 10, ΔCS_{IP} would be about \$90 million, in contrast to its baseline value of \$27 million.

7.3 Substitution Parameter σ

Each value of σ gives us a new vector of product qualities δ, which we term $\delta(\sigma)$. Each new δ vector, in turn, can be used to construct forecasts of expected quality. We can use these to create estimates of ΔCS_{IP} and the ΔCS ratio to see the sensitivity of these measures to σ. We would expect the absolute change in ΔCS_{IP} to depend on the substitution parameter σ, and Figure 1 confirms that this is so. Using our baseline σ of 0.5191, ΔCS_{IP} is \$27 million. By contrast, if σ were 0.25 or 0.8, then ΔCS_{IP} would be about \$45 million or \$13 million, respectively. (These estimates are visible on the vertical line above 0.33 in Figure 1).

It is not clear a priori how different levels of substitution affect the ΔCS ratio, so we undertake simulations for different values of σ. Figure 2 depicts the relationship between σ and the ΔCS ratio. Our estimate of the ratio is nearly invariant to our choice of σ. If $\sigma = 0$, then this becomes the plain logit model, and the ΔCS ratio is 19.78, while if $\sigma = 0.9$, the ratio is 19.88. Because σ is the only estimated parameter determining δ, Figure 2 also contains implicit estimates of the standard error of our ΔCS ratio estimate. That the ΔCS ratio is nearly invariant in σ means that if we take bootstrap draws from the estimated σ distribution, the resulting values of the ΔCS ratio would be tightly distributed. We conclude that our estimates of the ΔCS ratio are not sensitive to the choice of logit vs nested logit, nor are they sensitive to the degree of substitutability among songs. Beyond this, our estimates of the ΔCS ratio are precise.

7.4 Investors' Forecasting Ability

One of the key features of the model is the extent to which investors can forecast quality at the time of investment. The better their ability to forecast, the smaller are both ΔCS_{IP} and the ΔCS ratio. Hence, we would like to investigate the sensitivity of our welfare estimates to different abilities to forecast.

Strictly speaking, what matters for the estimated magnitudes of ΔCS_{IP} and the ΔCS ratio is not the R^2 from the forecasts per se but rather the value of the choice set the prediction model places in the counterfactual. Recall that $CS = \frac{M}{\alpha} \ln \left(1 + \sum_{j \in pred} e^{\frac{\delta_j}{1-\sigma}}\right)$, where $j \in pred$ refers to the set of products j predicted to be in the counterfactual choice set. What matters, therefore, for a forecast is $\sum_{j \in pred} e^{\frac{\delta_j}{1-\sigma}}$. Of course, R^2 and $\sum_{j \in pred} e^{\frac{\delta_j}{1-\sigma}}$ are related. To see this, note that with perfect prediction, and therefore R^2 of 1, the songs predicted to be in the top third are those actually appear in the top third, or, $\sum_{j \in pred} e^{\frac{\delta_j}{1-\sigma}} = \sum_{j \in actual} e^{\frac{\delta_j}{1-\sigma}}$.

Ideally, we would like to see beyond the veil of our ignorance to understand how our forecasting ability improves as we add more variables. Of course, we have already included all of the variables available to us in our forecast. To see how our estimate would change if we had better ability to forecast, we can create a new explanatory variable that is the true value of δ plus a scaled random error. That is, define $B_j = \delta_j + s\upsilon_j$, where s is a scaling variable which we control and υ is a standard normal error.

Then our forecasting model regresses δ on Z as in Section 5.2 above, along with B. We begin with a large value of s, so that we are adding an irrelevant variable, whose coefficient will be small.[26] As s shrinks, B acquires a significant coefficient; and our ability to predict quality improves. Each value of s is thus associated with a regression R^2 and a prediction R^2. Figure 3 depicts the relationships between ΔCS_{IP} and the ΔCS ratio and the prediction R^2.

When $s = 1000$, the regression R^2 rises from its baseline of 0.403 to 0.619; and the associated prediction R^2 (for vintage 2011 alone) rises from its baseline of 0.325 to 0.644. The ΔCS ratio would fall from its baseline value of nearly 20 to about 2.5, and the absolute change ΔCS_{IP} would fall from its baseline of \$27 million to about \$3.5 million. When $s = 300$, the

[26]We use the following values of s to perform our exercise: 10000000, 5000000, 1000000, 500000, 100000, 10000, 1000, 900, 800, 700, 600, 500, 450, 400, 350, 300, 250, 200, 150, 100, 50, 10, 5, 1, and 0. A value of $s = 10000000$ gives rise to our baseline prediction R^2 of 0.325.

regression R^2 rises to 0.929, and the associated prediction R^2 is 0.931.

A conservative approach to measuring the welfare gain from digitization would employ the richest and most accurate prediction model available. One model that errs on the side of conservatism is the model estimated on 2011 data so that the regression residuals are direct "forecasts" of quality (as opposed to using the forecasts derived from the 2010 regression). That regression had an R^2 of 0.411. The resulting estimates of ΔCS_{IP} and the ΔCS ratio are \$17.97 million and 13.04, respectively.

7.5 Alternative Demand Model

Bryniolfsson et al. (2003) estimate the benefits of the long tail by calculating the share of book sales accounted for by books available online but not likely to be available at consumers' local stores. In particular, following Hausman (1981), they estimate the change in consumer surplus as $\frac{-p\Delta q}{(1+\epsilon)}$, where q is the purchased quantity of the books newly available online, and p is the price per unit of these new products, and ϵ is the price elasticity of demand for the new product.

We can use this approach to estimate the welfare benefit of the change in products from digitization, relative to the conventional long tail benefit. To this end, define Δq_r as the difference between status quo track sales and the sales of products that would have existed under regime r. Hence, for example, $\Delta q_{PF} = (q_0 - q_{PF})$.[27] Then our ratio of interest is as follows:

$$\frac{\Delta CS_{IP}}{\Delta CS_{PF}} = \frac{\frac{-p\Delta q_{IP}}{(1+\epsilon)}}{\frac{-p\Delta q_{PF}}{(1+\epsilon)}} = \frac{\Delta q_{IP}}{\Delta q_{PF}}. \tag{15}$$

In short, this is sales of the products made available in the imperfect predictability simulation over the sales of products made available with perfect foresight. We calculate this to be 19.75. Note that this estimate is very similar to the one obtained from our baseline model, although the approach is vastly different.

[27]Note that this approach implicitly assumes that total counterfactual sales would equal the sales of the products predicted to exist in the counterfactual in the status quo (when they are available alongside the remainder the status quo products).

8 Conclusion

Our study has three conclusions. First, unpredictability can have a large effect on the impact of new product entry on welfare. Evaluating the benefit of new products is a central task for economics. We explore the welfare benefit arising from the new products prompted by reduced entry costs in a context in which quality is unpredictable. This unpredictability has a large effect on the benefit of new products. Given that unpredictability appears to be a common feature of new products, this idea may have wider applicability.

Second, applying this perspective to the impact of digitization on the recorded music industry yields some novel insights about the benefit of the Internet. Observers have understood the benefit of the Internet to operate through a shelf-space mechanism that we have termed the long tail in consumption. As important as this mechanism is, we propose that the long tail in production that we explore is quantitatively important. Reductions in entry costs allow producers to "take more draws," and given the unpredictability of quality at the time of investment, taking more draws can generate more "winners." Our estimates for music show that the production mechanism generates almost 20 times as much benefit as the consumption mechanism for an equal-sized increase in the number of products. The absolute size of the gains is substantial as well. In our baseline estimates, US consumers gain about $US 118 per year. Unpredictability is a generic feature of creative products such as books and movies, suggesting that the growth of new products in those categories may be producing large welfare benefits (Waldfogel, 2013a; Reimers and Waldfogel, 2013).

Finally, the results of this study provide evidence of an explicit mechanism by which the growth in new music products since Napster has raised the realized quality of music, as Waldfogel (2012) and Aguiar et al. (2014) have argued, despite the collapse of recorded music revenue.

References

Aguiar, L., N. Duch-Brown, and J. Waldfogel (2014): "Revenue, New Products, and the Evolution of Music Quality since Napster," *IPTS Working Paper*.

Anderson, C. (2006): *The Long Tail: Why the Future of Business Is Selling Less of More*, Hyperion.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841–90.

Berry, S. T. (1994): "Estimating Discrete-Choice Models of Product Differentiation," *RAND Journal of Economics*, 25, 242–262.

Berry, S. T. and J. Waldfogel (1999): "Free Entry and Social Inefficiency in Radio Broadcasting," *RAND Journal of Economics*, 30, 397–420.

Björnerstedt, J. and F. Verboven (2013): "Merger Simulation with Nested Logit Demand - Implementation using Stata," Konkurrensverket Working Paper Series in Law and Economics 2013:2, Konkurrensverket (Swedish Competition Authority).

Brynjolfsson, E., Y. J. 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.

Cardell, N. S. (1997): "Variance Components Structures for the Extreme-Value and Logistic Distributions with Application to Models of Heterogeneity," *Econometric Theory*, 13, 185–213.

Caves, R. (2000): *Creative Industries: Contracts Between Art and Commerce*, Harvard University Press.

Danaher, B., Y. Huang, M. D. Smith, and R. Telang (2014): "An Empirical Analysis of Digital Music Bundling Strategies," *Management Science*, 60, 1413–1433.

Ferreira, F., A. Petrin, and J. Waldfogel (2013): "Trade, Endogenous Quality, and Welfare in Motion Pictures," Working paper.

Goldman, W. (1984): *Adventures in the Screen Trade*, Grand Central Publishing: New York.

Gourville, J. (2005): "The Curse of Innovation: A Theory of Why Innovative New Products Fail in the Marketplace," Harvard Business School Working Paper 06-014.

Handke, C. (2012): "Digital copying and the supply of sound recordings," *Information Economics and Policy*, 24, 15 – 29, the Economics of Digital Media Markets.

Hausman, J. A. (1981): "Exact Consumer's Surplus and Deadweight Loss," *American Economic Review*, 71, 662–76.

Hausman, J. A. and G. K. Leonard (2002): "The Competitive Effects of a New Product Introduction: A Case Study," *The Journal of Industrial Economics*, 50, pp. 237–263.

Oberholzer-Gee, F. and K. Strumpf (2010): "File Sharing and Copyright," in *Innovation Policy and the Economy, Volume 10*, National Bureau of Economic Research, Inc, NBER Chapters, 19–55.

Petrin, A. (2002): "Quantifying the Benefits of New Products: The Case of the Minivan," *Journal of Political Economy*, 110, pp. 705–729.

Reimers, I. and J. Waldfogel (2013): "Storming the Gatekeepers: Digital Disintermediation in the Market for Books," *Working Paper*.

Shiller, B. and J. Waldfogel (2011): "Music for a Song: An Empirical Look at Uniform Pricing and Its Alternatives," *The Journal of Industrial Economics*, 59, 630–660.

Sinai, T. and J. Waldfogel (2004): "Geography and the Internet: is the Internet a substitute or a complement for cities?" *Journal of Urban Economics*, 56, 1–24.

Terviö, M. (2009): "Superstars and Mediocrities: Market Failure in the Discovery of Talent," *Review of Economic Studies*, 76, 829–850.

Verboven, F. (1996): "International Price Discrimination in the European Car Market," *RAND Journal of Economics*, 27, 240–268.

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

——— (2013a): "Cinematic Explosion: Movies, Gatekeepers, and Product Discovery in the Digital Era," *Working Paper*.

——— (2013b): "Digitization and the Quality of New Media Products: The Case of Music," in *Economics of Digitization*, University of Chicago Press.

A Figures and Tables

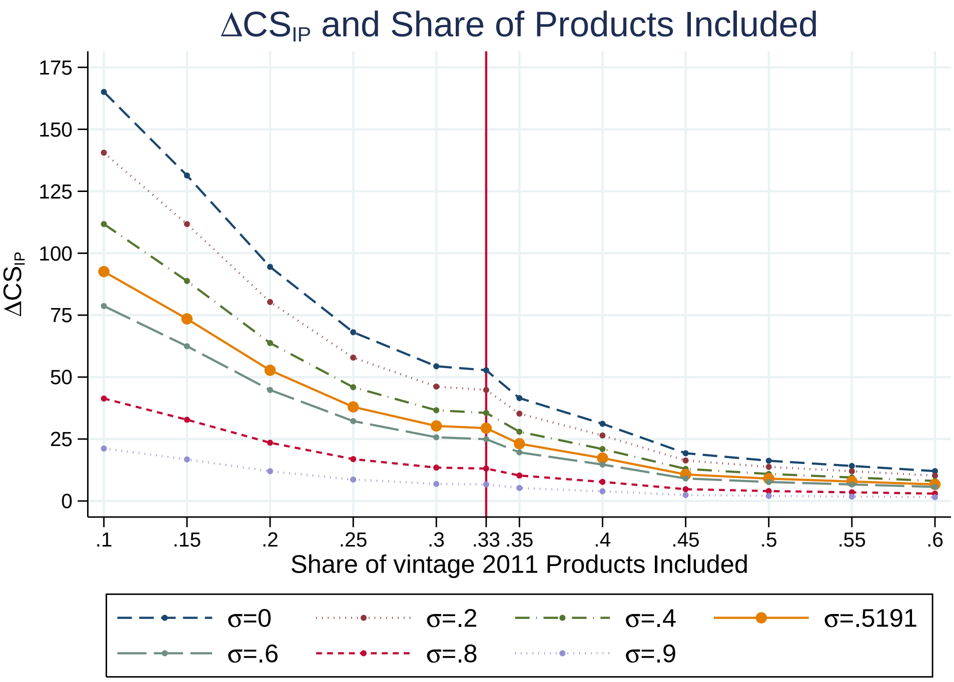

Figure 1: ΔCS_{IP}, σ, and Shares of Products Included.

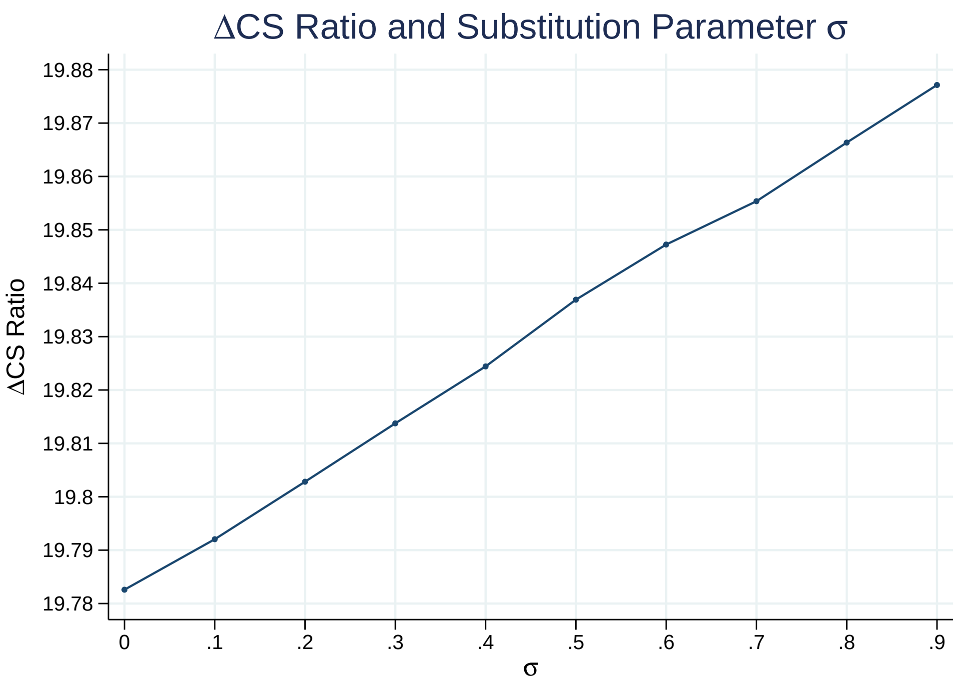

Figure 2: ΔCS Ratio and Substitution Parameter σ.

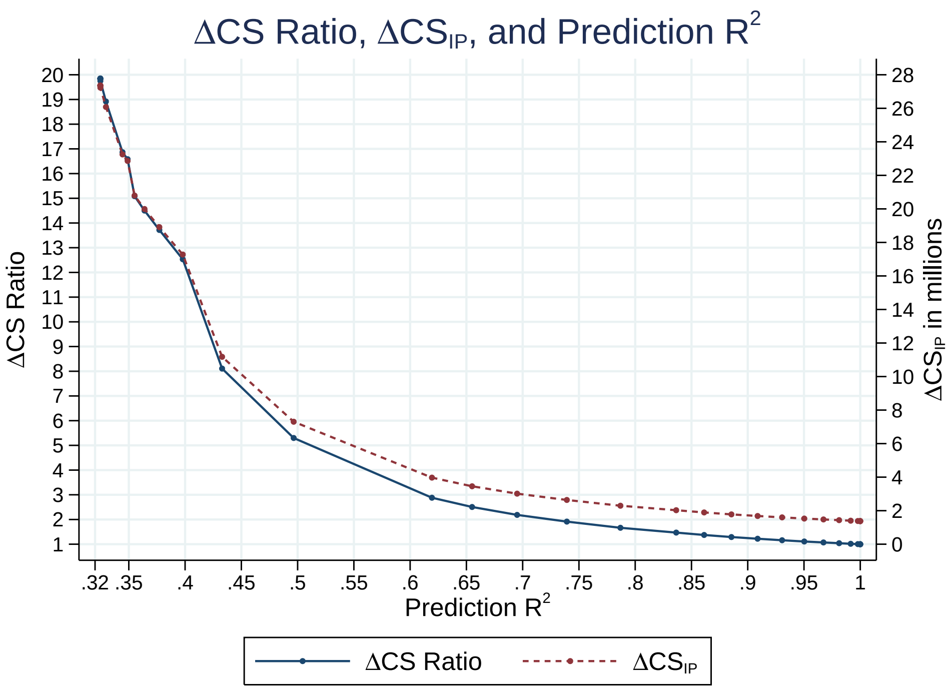

Figure 3: ΔCS Ratio and R^2.

Table 1: Demand Model

	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.	(5) Coef./s.e.	(6) Coef./s.e.
$\ln(s_j/s_0)$	0.858***	0.607***	0.536***	0.555***	0.519***	
	(0.03)	(0.07)	(0.06)	(0.06)	(0.05)	
$\ln(s_j/s_g)$						0.558***
						(0.03)
$\ln(s_g/s_0)$						0.506***
						(0.20)
Share of Digital Sales	6.172***	3.850***	3.298***	3.430***	3.109***	3.432***
	(0.48)	(0.71)	(0.66)	(0.67)	(0.65)	(0.37)
Age of the song	-0.015	0.259***	0.455***	0.444***	0.498***	0.427***
	(0.01)	(0.07)	(0.08)	(0.09)	(0.08)	(0.05)
Year Fixed Effects	✓	✓	✓	✓	✓	✓
Genre Fixed Effects	✓	✗	✓	✓	✓	✓
Origin Fixed Effects	✓	✓	✗	✓	✓	✓
Instruments (sums of)	-	Age	Age, Genre	Age, Origin	Age, Origin, Genre	Age, Origin, Age within broad genre, Origin within broad genre
R^2	0.829	0.759	0.719	0.732	0.708	0.733
No. of Obs.	2384157	2384157	2384157	2384157	2384157	2384157

† Specifications (1) uses OLS. Specification (2) - (5) correspond to the one-level nested logit model. Specification (6) correspond to the two-level nested logit model using genres as nests. Standard errors are clustered on country and year and are in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table 2: Forecasting Model

	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.
Log(sales in t-1)	185.868***	171.309***	126.021***	133.882***
	(1.62)	(1.63)	(1.51)	(1.71)
Log(sales in t-2)	-34.778***	-34.443***	-23.499***	-8.959***
	(1.90)	(1.89)	(1.71)	(2.00)
Log(sales in t-3)	-4.154**	-3.436*	-7.512***	-20.065***
	(1.84)	(1.82)	(1.66)	(2.01)
Log(sales in t-4)	14.658***	9.096***	0.078	-5.087**
	(1.39)	(1.38)	(1.27)	(1.98)
Log(sales in t-5)				-9.394***
				(1.49)
New Artist	1.274***	1.198***	0.876***	1.187***
	(0.01)	(0.01)	(0.01)	(0.01)
Artist's Age	-34.112***	-28.798***	-24.396***	-21.767***
	(0.70)	(0.71)	(0.65)	(0.73)
Artist's Age squared	0.376***	0.334***	0.294***	0.236***
	(0.01)	(0.01)	(0.01)	(0.01)
Years Since Last Release	27.508***	22.993***	16.645***	18.867***
	(1.20)	(1.19)	(1.08)	(1.35)
Origin Fixed Effects	✗	✓	✓	✓
Genre Fixed Effects	✗	✓	✓	✓
Label Fixed Effects	✗	✗	✓	✓
R^2	0.196	0.225	0.403	0.411
Prediction R^2	0.200	0.213	0.325	0.411
No. of Obs.	156411	156411	156411	134241

† Specifications (1) to (3) use 2010 data (t=2010) and songs from vintage 2010. Specification (4) uses 2011 data (t=2011) and songs from vintage 2011. The predicted δ's are constructed for the vintage 2011 songs in all specifications. The prediction R^2 is computed as the square of the correlation between the realized δ's in 2011 and their predictions. Standard errors are in parenthesis.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table 3: Fixed Costs of Entry.†

Regime	Perfect Foresight	Imperfect Predictability	No Predictability
Counterfactual	133.96	2515.76	10336.59
Status Quo	6.09	26.57	9467.89

† Fixed costs are estimated as the expected US digital single revenue of the last entering product, scaled up to the size of the entire US recorded music market in 2011. Status quo refers to the set of products available in the US in 2011, while counterfactual models the choice set if digitization had not occurred, referring to simulations in which the bottom two thirds of vintage-2011 products, by expected revenue, are removed from the choice set. Under perfect foresight, products are ordered by realized revenue. Under our main model, imperfect predictability, products are ordered by expected revenue. With the no prediction model, products are ordered randomly (so that the counterfactual choice set has one third of actual vintage-2011 products, chosen at random). All figures are in $US 2011.

Table 4: Counterfactual Results.[†]

Regime	ΔCS	Ratio CS	ΔRev	Ratio Rev	ΔTC.	Ratio TC	ΔW	Ratio W
Perfect Foresight	1.38	1	1.48	1	-5.18	1	8.03	1
Imperfect Predictability	27.35	19.84	29.34	19.86	-109.01	21.05	165.69	20.63
No Predictability	417.05	302.55	452.99	306.60	808.45	-156.15	61.59	7.67

[†] ΔCS is the change in CS from the tripling of the vintage-2011 products made possible by digitization. The three regimes differ by which products are in the counterfactual (no digitization) choice set. Perfect foresight adds products with the lowest realized quality, while imperfect predictability adds products with the lowest expected quality. The no predictability regime adds products that are as good, on average, as the products that would be available without digitization. "Ratio CS" reports ΔCS relative to the perfect foresight estimate that corresponds to the traditional long tail. ΔRev, ΔTC, ΔW, and the corresponding ratios are defined analogously. TC is the fixed cost per product times the number of entering products.

B Appendix

We would like a measure of the number of new products released each year. Unfortunately, our data cover only calendar years 2006-2011, so we have an index of the number of products first sold in each year for only 2006-2011. However, we can use our data to create a measure of the number of products released in each year. In each calendar year's sales data we see the number of products sold in that year originally released at each previous vintage. For example we could use the number of products from each vintage sold in 2011 as an index of the number of products released at each vintage. The only shortcoming of that approach is that some products from prior vintages will not show up in the sales data for each year; and just as sales tend to drop off over time, the probability of selling at least one copy may drop off.

A simple solution to this problem is to get a measure of the number of products from each vintage, controlling for age. To this end, define N_{tv} as the number of products from vintage v sold in year t, with age therefore given by $t - v$. Then we can run the following regression on the US data, 2006-2011:

$$\log(N_{tv}) = \theta_{t-v} + \gamma_v + \epsilon_{tv}, \tag{16}$$

where θ_{t-v} are flexible age effects, γ_v are vintage effects, and ϵ_{tv} is an idiosyncratic error. Then $\widehat{N_v} = e^{\gamma_v}$. Figure B.4 compares this index of the $\widehat{N_v}$ (implied new songs) with the number of songs for which the vintage equals the calendar year for 2006-2011. The figure also includes a horizontal line at the implied number of songs first released in 1999.

We are also interested in the share of sales occurring in each year of a song's life. Using our data for 2006-2011, we can directly observe the sales of a vintage-2006 song in 2006-2011, but this does not tell us how much of the sales occur after the sixth year. We can estimate the share of sales by age using an approach analogous to the approach above. That is, we can run the following regression: $\log(q_{tv}) = \theta_{t-v} + \gamma_t + \epsilon_{tv}$, where q_{tv} is the quantity of year-t sales that are for songs of vintage v, γ_t is a dummy for calendar year t, and other variables are as above. By exponentiating θ_{t-v} we get an index of the sales at each song age. In our data, sales decay with age. Sales shares for songs over 50 years old tend to be quite small.

We can accurately estimate the share of lifetime sales at age a as $\frac{s_a}{\sum_{a=0}^{80} s_a}$.

Using our data, the share of sales occurring in the first calendar year of release is 18.3 percent, followed by 20.6 percent in the second year, 9.4 percent in the third, 6.4 percent in the fourth, and 4.9 percent in the fifth.

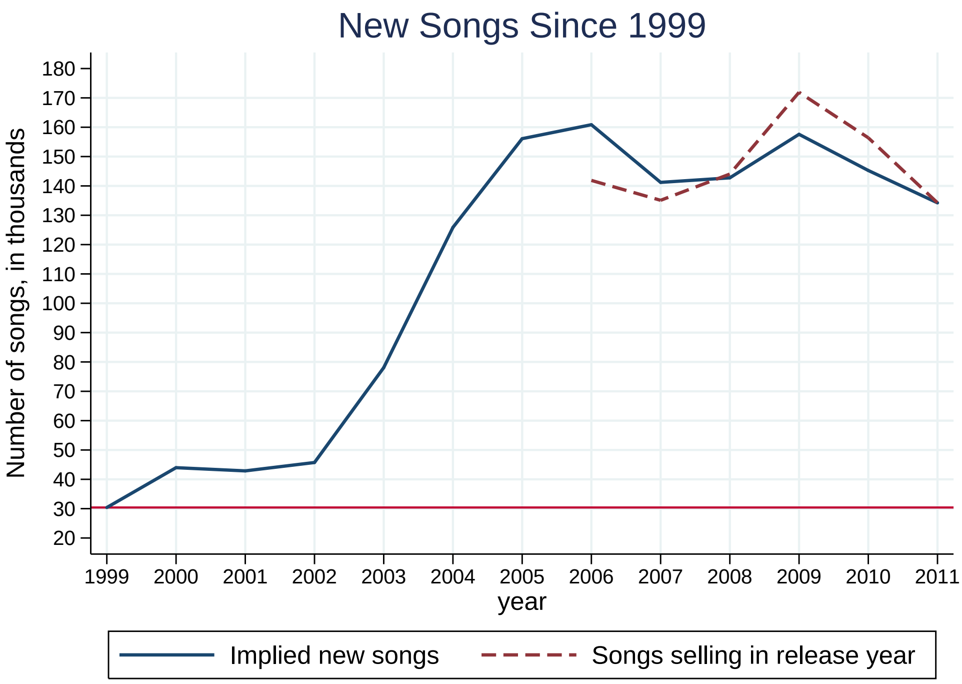

Figure B.4: New Songs Released, by Vintage.

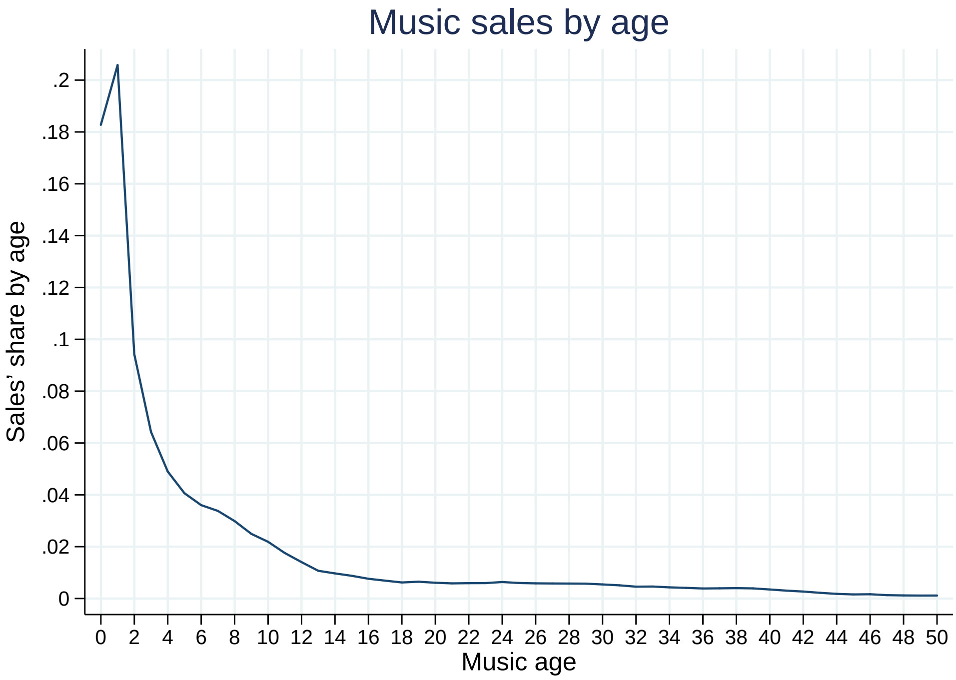

Figure B.5: Music Sales, by Age.