JRC/IPTS Digital Economy Working Paper 2014-XX

Panning for Gold:
The Random Long Tail in Music Production*

Luis Aguiar[†1] Joel Waldfogel[‡2]

[1]EC Joint Research Center - IPTS

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

July 2, 2014

Draft Version.
Please do not cite or circulate.

*Disclaimer: The views expressed are those of the authors and may not in any circumstances be regarded as stating an official position of the JRC or the European Commission.

[†]E-mail: luis.aguiar@ec.europa.eu

[‡]E-mail: jwaldfog@umn.edu

Abstract

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 such as books, movies, and music, then creating new products can have substantial benefits. Technological change in the recorded music industry tripled the number of new products between 2000 and 2010. 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 about ten times as much consumer surplus as the usual long-tail benefits from a tripling of the choice set according to realized quality. Our results also shed light on recent controversy over whether new products would deconcentrate consumption into a long tail or concentrate consumption toward "blockbusters": it depends on the new products' ex-post realizations.

Keywords: Digital Music, Digitization, Copyright, Entry.

JEL classification: L00, L82, O33.

1 Introduction

Since the early days of the Internet, researchers and others have extolled the virtue of the Internet's infinite shelf space, allowing consumers access to both popular and obscure products. And, indeed, the liberation of consumers from the tastes of their geographic neighbors is both interesting and important. 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). Benefits of the Internet to preference minorities may be particularly large (Sinai and Waldfogel, 2004). Despite the importance of long-tail effects in consumption, we will argue that they are even more important in production when we account for the unpredictability of product quality at the time of investment.

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. Industry observers say that roughly 10 percent of new movies are commercially successful, and the figures for books and music are similar (Caves, 2000). When the costs of bringing new products to market fall, society can "take more draws from an innovation urn." If the appeal of new products to consumers were 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.

Technological change in the recorded music industry has allowed substantial growth in the number of new products in the choice set. Between 2000 and 2010 the number of new

products brought to market annually tripled, leading us to ask how a tripling of the number of available products affects welfare. We can measure the benefit as the difference between welfare with the new, enlarged choice set and a smaller choice set consisting of a third of existing products. By definition, the incremental products had lower *expected* revenue than the types of products already coming to market (otherwise they would already have been brought to market). But the welfare impact of an entry cost reduction that triples the choice set depends heavily on *which* third of existing 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, 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 new entry 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 is not entirely predictable, 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 estimates of the realized appeal of each product (δ). We use the realized δ's 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 question that motivates the paper: 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 welfare? And 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 about 10 times as much consumer surplus and producer revenue than a tripling of the choice set according to realized quality.

The paper proceeds in 6 sections after the introduction. Section 2 starts by presenting descriptive facts about entry in the music industry 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 estimations, 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 object of interest, the welfare impact of an enlarged choice set with imperfect predictability of product appeal, in relation to the welfare impact of an enlarged choice set with perfect foresight. Section 7 discusses robustness of estimates. Section 8 concludes and discusses the policy implications of our results.

2 Background

2.1 Industry Background

Since 1999 recorded music revenue has fallen 70 percent around the world (see Figure 1). 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 Waldfogel (2013) argues, the cost of production, promotion, and distribution of new music have fallen sharply with digitization. And as Figure 2 shows, the number of new recorded music products brought to market each year has risen since 1990 and more sharply since 2000 (Aguiar et al., 2014). According to Nielsen data, the number of new music products brought to market tripled between 2000 and 2010. 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.

The idea that the Internet delivers more varied consumption opportunities has been explored widely. Brynjolfsson et al. (2003) quantify the benefit of access to the full list of books at Amazon in contrast to the 50,000 books available to a local 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 production mechanism that we explore.

More recent work raises questions about how the availability of additional products would affect sales concentration. Brynjolfsson et al. (2010) discusses the possible effects of greater variety on concentration measures. While the general of the literature suggests that the availability of more niche varieties would deconcentrate consumption over from the head, Elberse (2013) argues that the digitization era has seen a growth in "blockbusters," or products at the head of the distribution. She casts doubt on the importance of the long tail, with evidence of continued - and growing - concentration of sales in music and other products.

2.2 How Would Entry Cost Reduction Affect Welfare?

To fix ideas this section articulates a simple model of entry with heterogenous revenue and fixed entry costs. 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: $y' = y + \varepsilon$. 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 ($\varepsilon = 0$), 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 available products. 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 products. The major challenge to this exercise, however, is determining *which* third of status quo products would have existed if digitization had not reduced entry costs. This, in turn, depends on the predictability of product quality.

If product quality were completely predictable, then when costs were high, only the products with the highest realized quality would enter. Hence, the counterfactual high-entry-cost choice set is 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 the usual approach, the benefit of additional products would be relatively small. At the other extreme, if quality were completely unpredictable, 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.

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.

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 under any entry cost structure. The entry model infers fixed costs from the expected revenue of the last entering product.

3 The Model

This section describes the components of our model, demand quality prediction, and entry. Details of 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. Suppressing the time subscript, each consumer therefore decides in each month whether to download one song in the choice set $J_c = \{1, 2, 3, \ldots, J_c\}$ or to consume the outside good (not downloading 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. 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.[1] The parameter ζ_i therefore 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

[1] 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.

to the simple logit (see footnote 1) 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_j.

3.2 Quality Prediction

The results from our demand estimation allow us to construct estimates of the mean utility of each song following equation (12). While demand estimation uses data for multiple countries, we undertake the entry exercise 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 measures how much appeal it generates to consumers once the song has entered the market. 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 artist a 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):

$$\delta_{ja} = \gamma_0 + \gamma_1 Z_a + \upsilon_{ja}, \tag{2}$$

where the vector Z_a contains information on artist a and υ_{ja} is an error term. The predicted values $\delta'_{ja} = \widehat{\gamma}_0 + \widehat{\gamma}_1 Z_a$ then provide us with a measure of the expected quality of each song j prior to release.

3.3 Supply and Fixed Costs

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, 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, 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. The minimum revenue observed in our data is 1 US sale in 2011 at a price of \$1, so the implied status quo fixed cost is \$1 as well. To estimate fixed costs that give rise to one third of status quo entry, we must calculate the expected revenue of the last product when only the best-selling products ($\frac{N}{3}$) enter.

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 function of the vector Δ_k.

If we order the products such that $\delta_k > \delta_{k+1}$, the products enter as long as $\mathrm{E}[r_k(\Delta_k)] > T$. For example, given the nested logit structure, the expected and realized revenue to product 1 when it is alone is

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

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

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

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

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: $T \approx r_k, \quad k = \frac{N}{3}$.

3.3.2 No Predictability

At the opposite extreme from the perfect predictability model is a model with no predictability. With no predictability, all products are identical ex-ante. Hence the expected revenue of any product depends only on the total number of products entering (n) and is the total revenue to those n products divided by n. That is,

$$\mathrm{E}\left[r_n\right] = pM\mathrm{E}\left[\frac{\frac{D_n}{D_n^\sigma + D_n}}{n}\right], \tag{5}$$

where D_n is evaluated with a particular draw of n product qualities (δ_i), 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 prediction 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: $\delta'_1 > \delta'_2 > \ldots > \delta'_K$. 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'). Then calculate the realized revenue to each song using the realized qualities in the logit formula. Thus, the model's expected revenue for the k^{th} entrant, when it is last, is given by a formula that resembles (4), except that the realized δ's are ordered by predicted δ (δ', or expected quality):

$$\mathrm{E}\left[r_k\right] = pM\mathrm{E}\left[\frac{e^{\frac{\delta_k}{1-\sigma}}}{D_k^{\sigma} + D_k}\right] \tag{6}$$

This approach has the feature that the realized revenue for the k^{th} product is noisy. Prominent in the share formula is the $e^{\frac{\delta_k}{1-\sigma}}$ of the numerator. Precisely because quality is difficult to predict, the particular realization for the k^{th} product fluctuates across entrants k. These fluctuations are independent across k, so we can characterize the function relating the number of products entering to the expected revenue of the last product entering by smoothing the expected revenue function.

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 product when the top $\frac{N}{3}$ products according to expected quality enter.

3.3.4 Welfare Counterfactual

Each of the three approaches provides both a set of entering products and a method for calculating the expected revenue of the last entering product and therefore an estimate of fixed costs. Given an estimate of fixed costs (T), along with an approach to calculating consumer surplus and revenue associated with a number of entering products, we can calculate the welfare associated with any level of entry n as $W(n) = CS(n) + rev(n) - nT$. We can assess the impact of technological change on welfare by comparing an evaluation of the function at status quo n and T against its value at pre-digitization n_0 and T_0, where the 0 subscripts refer to the counterfactual with one third of the status quo products.

4 Data

The data used in our study come from Nielsen and include the digital sales of recorded music in the US, Canada, and 15 major European countries between 2006 and 2011.[2] We observe the annual sales of each downloaded track in each destination. The dataset originally includes 1,532,095 distinct artists but unfortunately does not include the artists' country of origin. To overcome this shortcoming of the data, we recovered data on artists' country of origin from www.musicbrainz.org, an open music encyclopedia that collects music metadata and makes it available to the public.[3] The MusicBrainz database is sufficiently authoritative that the BBC relies on it to support the artist and music information on their music website.[4] Unfortunately, there is no unique identifier that permits a straightforward matching between our data and MusicBrainz. We therefore engaged in a tedious matching procedure based on artists' names. In order to reduce the burden of the matching procedure while still obtaining a sufficiently representative set of artists, we decided to focus on the matching of the top 150,000 selling artists in Nielsen, which accounts for 99% of the overall sales. We could obviously not assign a country of origin to illegible artists or to observations that correspond to non standard artists such as movie soundtracks or compilations. On top of that, not all artists appearing in the Nielsen data were present in MusicBrainz, which unfortunately

[2] 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.

[3] Whenever available in MusicBrainz, the country of origin of each artists corresponds to their country of birth.

[4] See http://www.bbc.co.uk/music/brainz/.

prevented us from recovering the origin of all the artists in our data.[5]

After excluding observations to which no origin country could be assigned, we end up with a sample of 75,239 distinct artists that cover over 91% of the Nielsen total sales. While many artists remain unmatched, it is worth mentioning that over 5,600 of artists sell fewer than 500 copies in the sample, meaning that we are still left with a significant long tail. Our data include 3,987,877 distinct tracks and, because a song can appear in multiple countries and years, 50,870,037 observations. Total track sales in the data are 628.3 million in 2006 and rise to 1512.4 million in 2011. See Table 2.

Digital music markets have developed to different extents across countries. Table 1 provides a comparison of the various countries' digital sales in the Nielsen data, their populations, and total music as well as digital music sales from IFPI (2013). As of 2011 digital sales made up 56 percent of total music sales in the US, compared with 22 percent in France.

5 Empirical Implementation

5.1 Demand Model

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

$$\begin{aligned} \ln(S_{jc}) - \ln(S_{0c}) &= \delta_{jc} + \sigma \ln\left(\frac{S_{jc}}{1-S_{0c}}\right) \\ &= x_{jc}\beta - \alpha p_{jc} + \sigma \ln\left(\frac{S_{jc}}{1-S_{0c}}\right) + \xi_{jc}, \end{aligned} \tag{7}$$

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.

[5] We faced an additional complication in that not all artists matched with MusicBrainz had a country of origin assigned. To partially solve this problem, we looked for these specific artists in Wikipedia in order to recover their country of origin.

Intuitively, σ depends on how the total inside share of songs (i.e. the total share of the entire digital market for downloading of music) changes as the number of songs in the choice set varies. When σ is close to one, the total inside share does not vary much with the number of songs since in this case the within group substitution is high. In other words, additional songs will simply cannibalize other songs' shares since consumers will substitute old songs for new songs (business stealing). At the opposite extreme is the case where $\sigma = 0$. In this case adding new songs to the choice set will lead some consumers of the outside good to substitute to a new song when it is added to the choice set. In this case the total inside share of songs will therefore increase.

5.1.1 Identification of σ

Since the inside share of each song j is, by construction, endogenous in equation (7), we need to find an instrument in order to consistently estimate σ. Given that identification of σ is related to how the inside shares change and the number of songs available in a market change, and since we observe variation in the number of songs available over time and across markets, one potential instrument is J_c, the number of songs available in a given country. Figure 3 graphs the relationship between the number of products (songs) and the total inside share across countries for the year 2011. For each country c, the total inside share is defined as $S_{J_c} = \frac{1}{M_c} \sum_{j \in J_c} q_{jc}$, where M_c is a measure of market size and q_{jc} is the number of downloads for song j in country c.[6] Figure 3 shows a positive relationship between the number of songs available in a given country and the share of the population consuming music in the form of digital downloads. There is, however, reason for concern that the number of songs entering each market is endogenous. In particular, we would expect entry to be larger in markets with greater unobserved demand for music. An alternative instrument is country population. If more songs are made available in larger markets simply because of market size and not because of demand intensity, consumers in larger markets would face larger choice sets of music. Figure 4 presents the relationship between the number of songs available and the population of each country for the year 2011 and shows that larger countries do indeed offer larger choice sets. The descriptive evidence from figures 3 and 4 therefore suggest a positive relationship between the number of available downloadable songs and the total share of population consuming music in the form of digital downloads. This suggests a substantial

[6] For each country c, we define the market size as 12 times its population.

benefit from additional songs (market expansion) and therefore an estimate of σ lower than 1.

5.1.2 Consumer Surplus and Revenue

Given our estimates of σ and α, we can calculate the mean utility of each song, and given these estimates of δ_{cj} we can calculate the consumer surplus (CS) and revenue. The formula for the CS is given by

$$CS = \frac{M}{\alpha} \ln \left(\sum_J D_J^{1-\sigma} \right) = \frac{M}{\alpha} \ln \left(D_J^{1-\sigma} + 1 \right). \tag{8}$$

Revenue is given by

$$Rev = p_j M \left[\frac{D_J}{(D_J^{\sigma} + D_J)} \right]. \tag{9}$$

5.1.3 Price coefficient

We don't observe variation in price, but if we knew the average price as well as the marginal cost, then we could infer α from a profit maximizing assumption, as is customary in the literature. We treat the US price as \$1 per song, and we treat zero as the marginal cost.[7]

Notice that when we have only 2 nests (as in our case), then the music-elasticity $\eta_{jk} = \eta_{j0}$ since the good outside the nest is the outside good. This leads to:

$$\eta_{j0} = \alpha p_{0c} S_{0c} = \alpha p_{0c} \left[1 - \frac{D_{J_c}}{D_{J_c} + D_{J_c}^{\sigma}} \right] \tag{10}$$

Because the price is constant, the term αp_{jc} in (7) simply becomes part of the constant term.

[7] While Apple pays \$.70 for each song in the US, it is far from clear that they price in a double marginalized way. Shiller and Waldfogel (2011) find a revenue maximizing uniform song price that is close to the actual iTunes song price, suggesting that zero marginal cost is a reasonable assumption.

We therefore start by estimating

$$\ln\left(S_{jc}\right) - \ln\left(S_0\right) = x_{jc}\beta + \sigma \ln\left(\frac{S_{jc}}{1 - S_{0c}}\right) + \xi_{jc}, \tag{11}$$

where x_{jc} includes a constant and a set of year dummy variables. We get an estimate for σ, $\widehat{\sigma}$, that we can use to calculate the country-specific mean utility of each song $\widehat{\delta}_{jc}$:

$$\delta_{jc} = \ln\left(S_{jc}\right) - \ln\left(S_{0c}\right) - \sigma \ln\left(\frac{S_{jc}}{1 - S_{0c}}\right). \tag{12}$$

We infer α by assuming that pricing is governed by a revenue maximizing monopolist.[8] Then, assuming that $p = 1$ for all songs in all countries, we have that the elasticity of music is given by

$$\eta_{j0} = \alpha \left[1 - \frac{D_{J_c}}{D_{J_c} + D_{J_c}^{\sigma}}\right]. \tag{13}$$

Solving for the α that makes the demand for music unit elastic, we therefore obtain

$$\alpha = \frac{1}{\left[1 - \frac{D_{J_c}}{D_{J_c} + D_{J_c}^{\sigma}}\right]}, \tag{14}$$

where $D_{J_c} = \sum_{j \in J_c} e^{\frac{\delta_{jc}}{1-\sigma}}$. At this point we therefore have estimates of σ, α and mean utilities for traded songs j in destination countries c (δ_{jc}), which allow us to calculate CS and PS.

5.1.4 Results

The key parameter that we estimate is σ which, intuitively, is identified from the relationship between the number of products and the share of the population consuming. The expansion of the digital markets which proceeds at different rates in different countries produces a threat to identification. The growth in the number of digital products and digital consumption both

[8] 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.

arise from the diffusion of digital technology and may give the appearance that a growing number of products expands the market, or that σ is low. We employ two strategies to avoid this sort of mistaken inference. First, we instrument the inside share with cross sectional variation in measures of market size (population or its logarithm) for the destination country. Second, we employ direct and indirect controls for digital diffusion. We include a direct measure of digital share of music expenditure in each destination and year. We also include a host of other country level controls, including 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.

Table 3 reports estimates of the demand model when using the natural logarithm of population as the instrumental variable for the inside share in equation (7). Column (1) includes GDP per capita as well as the share of digital music expenditure (our measure of digital diffusion). On top of these control variables, specification (2) uses the logarithm of population interacted with time dummies as instruments for the inside share. Specification (3) builds on (2) and adds interactions between the control variables and time dummies. Finally, specifications (4) to (6) include the same explanatory variables and instruments as, respectively, specifications (1) to (3) and add other country-specific control variables: 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. We find estimates of σ ranging between .67 and .8. We also estimate the same specifications with different instrumental variables for the inside share, using the country population in levels, the number of products in the market, and the logarithm of the latter. Table 4 presents the estimates of σ for each specification and each set of instruments. The second row of the table therefore corresponds to the estimates already presented in the first row of Table 3. Overall we find estimates of σ ranging between .4 and .8.

Our preferred specification uses the logarithm of population interacted with time dummies as instruments for the inside share and includes all the control variables interacted with time dummies (specification (6) in Table 3). This specification gives the highest estimate of σ and therefore the most conservative estimate of the gains from trade liberalization. We will use these demand estimates for our counterfactuals, noting that the estimates from other specifications are very similar.

5.2 Quality Prediction

While we estimate the demand model on data for 17 countries over the period 2006-2011, we will perform our counterfactual exercises on US data for 2011. Hence, we require a quality forecasting model only for the 2.2 million songs sold in the US in 2011.

Our main interest is not in the level of welfare under the counterfactual or status quo regimes, nor in the difference in welfare between the status quo and counterfactual regimes per se, but rather in the difference between these changes under perfect foresight and imperfect predictability, $\frac{\Delta CS_{IP}}{\Delta CS_{PF}}$, where ΔCS_{IP} is the change in welfare with the expanded choice set under imperfect prediction, and ΔCS_{PF} is the analogous change under perfect foresight. We term this expression the ΔCS ratio. The simulations proceed by comparing a "status quo choice set," such as all of the songs available for sale in 2011, to a choice set that would have been available if entry costs had not fallen, such as the top third of products according to expected quality at the time they were originally released.

Constructing counterfactual choice sets require a way of ordering products according to their prospects at release. It is tempting to regress δ_j for the 2011 releases on their characteristics at release, that is, the X_j's for the 2011 songs (including artists' past sales, etc.) and then to use the fitted values of δ_j as our measure of expected quality. This approach faces a few complications. First, a regression of the 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 $\delta_j' = X_{j,2011}\widehat{\beta}_{2011}$, the $\hat{\beta}$ from the 2011 regression contains the realizations for 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 $\widehat{\beta}$ from a regression of quality realizations for 2010 δ_j on $X_{j,2010}$, the resulting $\widehat{\beta}_{2010}$ can be applied to characteristics of 2011 songs to produce a prediction of the 2011 songs' qualities in 2011: $\delta_j' = X_{j,2011}\widehat{\beta}_{2010}$ that only uses information available prior to release.

A second complication is that the songs available in 2011 include both new songs first released in 2011 as well as older songs first released in prior years but still being sold. If songs depreciate over time, then δ_j will be lower for older songs because of their age, and not because they had lower expected quality at initial release. If we treat the top third of δ_j's as the counterfactual choice set, we might systematically exclude older songs, rather than songs of lower expected quality at release. A solution to this possible problem is to use the

top third of songs by realized quality in 2011 δ_j, by release vintage rather than overall, as the counterfactual choice set.

Given that our sales data begin in 2006 - and that we need prior year sales to make predictions - the earliest vintage for which we can predict quality using past sales is the 2008 vintage (when it's 3 years old, in 2011). This has the consequence that we are well positioned to model expected quality for the 2011, 2010, 2009, and 2008 vintages based on their artists' pre-release sales (as X's), along with their sales realizations in 2010 (determining which of the pre-2008 vintage songs would have been released if costs had been higher would require a different approach).

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. For artists who are not new, we observe past sales of their previous songs, which may be predictive to 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 sort themselves onto 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.

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 5, 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 and time since last release.[9] 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. Finally, label fixed effects are important. Their

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

addition, in column 2, raises the year-2010 regression R^2 from 0.196 to 0.386. Column 3 adds a host of interactions, with the goal of flexibly explaining more variation. These include the interaction of artist age and its square with past sales and time since last release, as well as the interaction of time since last release with past sales. The inclusion of these interactions raises the regression R^2 to 0.390.

The regression R^2 using data on 2010 δ is interesting, but it is not a direct measure of forecasting success in 2011. To see the fit of the forecast, we apply the coefficients in column 3 to the releases for 2011 (to predict the realized success of vintage-2011 songs in 2011). We can calculate a resulting "prediction R^2" as the square of the correlation of realized δ's with their predictions. The resulting R^2 measures, for regressions in columns (1-3) are 0.2, 0.315, and 0.316, respectively.

The last column of Table 5 provides a model for forecasting the success of songs in the year they are released. The 2.2 million songs available in the US in 2011 include 134,241 songs released in 2011, as well as another 194,419 songs originally released in 2010. To forecast the success of the 2010 releases in 2011, we need a forecasting model for predicting the success of vintage-2010 songs when they were one year old, using data available prior to their release. To this end, we estimate a model analogous to column 3 of Table 5, using data on vintage-2009 releases in 2010. When the resulting coefficients are applied to year-2010 releases, we can predict their δ's for 2011. We produce analogous forecasting models going back to predicting the success of vintage-2007 songs in 2010 (which we use to forecast the success of vintage-2008 songs in 2011). These regressions are reported in Table 6. The regression R^2 range from 0.390 to 0.456, and the R^2 associated with forecasting 2011 sales of the respective vintages range from 0.316 to 0.381. When we stack the resulting forecasts of 2011 delta for songs from vintages 2008-2011, the prediction R^2 - the squared correlation of realized 2011 delta and the forecasts - is 0.355.

Keeping in mind our interest in the relative change in welfare (e.g. $\frac{\Delta CS_{IP}}{\Delta CS_{PF}}$), we can entertain a variety of counterfactual choice sets to see how the estimate of our ratio of interest varies.

One simple comparison is between a status quo choice set consisting of the full 2011 choice set along with a counterfactual including all songs available in 2011 from long-past vintages while omitting the bottom two thirds of products, according to expected quality, from recent vintages (such as 2011, or the range 2008-2011). These simulations can be interpreted to

represent a cost reduction that occurred starting in, say, 2008 or 2011. Of course, the absolute change in welfare arising from the cost reduction that expands the choice set to status quo proportions will differ according to whether the simulated cost reduction arrived for the 2008 or the 2011 vintage. But it is worth noting, again, that we are interested in the relative change in welfare (e.g. $\frac{\Delta CS_{IP}}{\Delta CS_{PF}}$) and not in the absolute change.

Because of the importance of forecast quality for our results, we explore the sensitivity of results to different assumptions about forecast quality in Section 7.

5.3 Fixed Costs

Our estimates of fixed costs are based on estimates of the expected revenue of the last entering product. We estimate these in two different ways, treating different sets of products as "endogenous." First, we treat only the vintage-2011 products as endogenous, terming this the "2011 approach." That is, we treat the pre-2011 products available in 2011 as exogenously available. We then 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 data for a vintage 2011 song is \$1, the resulting status quo PF fixed cost estimate is \$1 (see Table 7). 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 \$22.01.

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 \$1555.28. 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 \$1740.05.

We calculate the imperfect predictability status quo fixed costs using the 2011 approach by ordering the 2011 products by expected quality. We then seek an estimate 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. The unpredictability of quality gives rise to large variation in the realized revenue of the k^{th} entering product, so we need to smooth. In particular, we average the realized revenues of the last 200 entering products, producing an estimate of \$44.64. For the counterfactual fixed costs under imperfect predictability, we average expected revenues over the 200 songs surrounding the song at the $(\frac{1}{3})$ percentile, producing a counterfactual fixed cost estimate of \$203.85.

As is customary in the empirical entry literature, our fixed cost estimates are derived from a cross section of revenue data. The fixed costs produced via the 2011 approach are estimated from 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. Given that song revenues tend to decline with the age of songs, we obtain different fixed cost estimates if we also include songs of earlier vintages.

We also estimate fixed costs treating all of the songs for which we have quality forecasts (vintages 2008-2011) as endogenous. We term this the "2008-2011 approach." The no predictability case is the most straightforward, so we describe it first. Status quo fixed costs are calculated as the average revenue to songs of vintage 2008-2011 in the year 2011, when they are marketed alongside songs with vintages before 2008. We calculate this as \$858.13. Not surprisingly, this is below the fixed cost implied by the 2011 approach, since that approach is based on first-year revenue. The no predictability counterfactual fixed costs are calculated as the average revenue to a randomly chosen third of songs of vintages 2008-2011 (marketed alongside the older songs). We estimate this as \$1280.66.

Under the 2008-2011 approach, we estimate the remaining fixed costs as the weighted average of vintage-specific fixed costs estimates. The example of counterfactual perfect foresight fixed costs provides a useful illustration. We can order the vintage 2008-2011 songs by quality within vintage. Under the counterfactual, only the top third of products from each of these vintages enters. We estimate the fixed costs associated with vintage 2008 songs in 2011 as the revenue of the vintage-2011 song at the 33^{rd}-highest percentile of quality, when the remainder of the choice set includes all pre-2008 songs as well as the top third of songs from vintages 2008-2011. We estimate analogous fixed costs associated with vintages 2009-2011, then average them according to the relative numbers of songs from vintages 2008-2011, respectively.

Using this approach we estimate status quo perfect foresight fixed costs as \$1. This is as before because the last entering product from each vintage 2008-2011 in 2011 has one sale. Counterfactual perfect foresight fixed costs are \$17.91. Imperfect predictability fixed costs under the status quo are \$15.47, while counterfactual imperfect predictability fixed costs under the 2008-2011 approach are \$224.80.[10]

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 5 to 14, or by about one order 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

Our main goal in this paper is comparing the long tail in production with the long tail in consumption. For example, Brynjolfsson et al. (2003) quantify the long tail benefits by comparing the consumer surplus available from the top 100,000 books to the consumer surplus delivered by access to all extant books. This exercise corresponds to our estimates of the benefits of the expanded choice set under perfect foresight. Table 8 reports our estimates of the welfare consequences of the cost reduction that raises the number of products in the choice set based on three counterfactual thought experiments: the tripling of products from all vintages, from vintages since 2008, and just for the vintage 2011.

The degree of predictability matters substantially for the result. If producers had perfect foresight, the gains from expanding the choice set - the conventional long tail benefits - would be small. Adding just more vintage-2011 songs raises consumer surplus by \$0.06 million, and adding more vintage 2008-2011 songs raises consumer surplus by \$0.30 million. At the other extreme - with no predictability - the analogous counterfactual expansions of the choice

[10] Fixed costs using the 2008-2011 approach should be lower than using the 2011 approach under imperfect predictability. This is violated in the estimates because the average of realized revenue does not decrease monotonically with expected quality we the bandwidth we have chosen.

set raises consumer surplus by 200 to 300 times more. That is, if it were literally true that nobody knew anything about which products would succeed, then the long tail in production would produce benefits 2-3 orders of magnitude larger than the benefits of the long tail in consumption. Results for revenue are similar.

Using our preferred imperfect predictability approach, expanding the choice set using just vintage-2011 songs raises consumer surplus by 16.82 times as much as the increase under perfect foresight. Using vintages 2008-2011, the increase is 11.16 times as much. Revenue results are quite similar.

This is the main result of our paper: with a degree of predictability that is plausible for an innovative process understood to be uncertain ("nobody knows anything"), the gain from a cost reduction that raises the number of products tested in the market is an order of magnitude higher than the conventional long tail benefit of an expanded choice set.

While our main focus is on the difference between ΔCS under imperfect predictability and perfect foresight, a few comments on the absolute size of our estimates are in order. First, the absolute size of ΔCS using the 2008-2011 approach, at almost \$0.3 million under perfect foresight and \$3.33 million under imperfect predictability, is small compared with the \$1 billion reported by Brynjolfsson et al. (2003) for books. Using our perfect foresight approach, the welfare benefits of the full choice set relative to the top 100,000 products is \$10.45 million.

We have focused thus far on the change in CS, but we can also examine the impact on full welfare, which includes the changes in CS, revenue, and fixed costs. As Table 8 shows, in absolute terms, most of the welfare gains come from the implied reduction in fixed costs. We estimated fixed costs from the expected revenue of the last entrant. We are hesitant to rely on our fixed cost estimates for measuring overall welfare. It seems likely that inframarginal entrants have higher fixed costs, in which case our approach - calculating total cost as $N \cdot FC$ - will underestimate the overall costs. Moreover, it is possible that the fixed costs inferred from the expected revenue of marginal entrants will imply a substantial reduction in total costs from digitization. We can view our fixed costs estimates $\mathrm{E}[r(N)]$ as a lower bound on the fixed cost. The upper bound is the expected revenue of a k^{th} entrant when all N products enter: $\mathrm{E}[r(k,N)]$. Hence, the lower bound of fixed costs is $n\mathrm{E}[r(N)]$, while the upper bound is $\sum_{k=1}^{N} \mathrm{E}[r(k,N)]$. The latter sum is total revenue, so if total costs equal total revenue, then welfare is consumer surplus.

7 Robustness

Our estimate of the ΔCS ratio depends on a host of underlying parameters, including the substitutability of products in the demand model (σ), the ability of investors to forecast quality at the time of investment (R^2), 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.

Our baseline counterfactual is a world in which all old - and only one third of recent - products exist. It is useful to know how the ratio of interest varies for different counterfactual shares (besides one third). To this end, we redo the counterfactual involving 2011 alone, including different shares of products between 0 and 1. Figure 5 depicts the ΔCS ratio ($\frac{\Delta CS_{IP}}{\Delta CS_{PF}}$) as a function of the share of 2011 products included in the counterfactual. By construction, the ratio is 1 when all products are included (when the share is 0, indicating that all products are excluded from the counterfactual). As the share rises, the ratio rises substantially, reaching 10 at a share of 0.1. The ratio is 16.82 at a share of one third (when the top third of products are included in the counterfactual). The ratio continues to rise as the share rises. We conclude that the random long rail is substantially larger than the conventional long for a wide range of choice set enlargements.

One of the important determinants of the sensitivity of welfare to the size of the choice set is the substitution parameter σ. Our baseline simulations use a high value of σ (0.817), which is conservative in the sense that it implies small welfare improvement with increases in the choice set. Of course, the result of interest is not the change in welfare per se but the ratio of changes in welfare under different expectations regimes. It is not clear (to us, anyway) a priori how different levels of substitution affect this ratio, so we undertake simulations for different values of σ. Recall that our estimates ranged between about 0.4 and 0.8. We choose various values of σ. Each value of σ gives us a new vector of product qualities δ, which we term $\delta(\sigma)$. We then use the new δ vector to create forecasts of expected quality. We use these to create estimates of the change in CS under imperfect predictability and perfect foresight. For these simulations we treat only 2011 songs as endogenous. Figure 6 depicts the relationship between σ and the ΔCS ratio. Our estimate of the ratio is nearly invariant to our choice of σ. We find that the ΔCS ratio remains high for a wide range of

σ values.

Because σ is the only estimated parameter determining δ, Figure 6 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 clustered. There is one complication. Our baseline estimate of σ is 0.817 (with a standard error of 0.11). Draws from $N(0.817, 0.11)$ yield some realizations with $\sigma \geq 1$, which gives rise to a discontinuous jump in the ΔCS ratio. Putting aside this possibility (which happens 9 times in 100 draws using baseline σ), we obtain a bootstrap standard error of the ΔCS ratio of 0.024. Note further that most of our estimates of σ in Table 3 have confidence intervals that are interior to $[0, 1]$ giving rise to a narrow range of ΔCS ratio estimates. We conclude that our estimates of the ΔCS ratios are precise.

We also explored the sensitivity of our estimates to the songs included in the choice set. Our baseline includes songs of all vintages in the status quo choice set, but we also calculated the ΔCS ratio including only vintage 2011 songs in the status quo choice set. Changing welfare from adding products has large effects on the *level* of CS but has little impact on the ΔCS ratio. This can be seen in Figures 5 & 6, where the red lines labeled "Excluding Exogenous Songs" represent the case where only vintage 2011 songs are included to compute ΔCS ratio.

One of the key features of the model is the extent to which investors can forecast quality at the time of investment, and we would like to investigate the sensitivity of our estimate to different abilities to forecast. In particular, the better their ability to forecast, the smaller the welfare gain under imperfect predictability relative to perfect foresight. Ideally, we would like to see beyond the veil of our ignorance to see 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 create a new explanatory variable that is the true value of δ plus a scaled random error. That is, define $Z_j = \delta_j + s\varepsilon_j$, where s is a scaling variable which we control and ε is a standard normal error.

Then our forecasting model regresses δ on X as in Section 5.2 above, along with Z. We begin with a large value of s, so that we are adding an irrelevant variable, whose coefficient will be

small.[11] As s shrinks, Z 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, as in Section 5.2) as well as a different value of the change in CS under imperfect predictability. When $s = 1000$, the regression R^2 rises from its baseline of 0.39 to 0.44; and the associated prediction R^2 (for vintage 2011 alone) rises from its baseline of 0.32 to 0.39. When $s = 300$, the regression R^2 rises to 0.71, and the associated prediction R^2 is 0.70. Using just the 2011 songs as endogenous, Figure 7 depicts the relationship between the ΔCS ratio and the regression R^2. As R^2 rises from its baseline value of 0.39 to 0.5, our estimate of the ΔCS ratio falls from 16.82 to about 8. When R^2 reaches 0.6, the ΔCS ratio falls below 5. When R^2 reaches 0.8, the ΔCS ratio falls below 2.

Of course we do not now know the true level of predictability of quality. We do know, however, that cultural industries are understood to be contexts in which quality predictability is challenging. To the extent that predictability is similar to what we characterize with our baseline model, the long tail in production will be much larger than the traditional long tail.

8 Conclusion

The long tail plays a central role in our understanding of the effects of digitization, as access to more products has expanded choice sets and raised ensuing consumer surplus. As important as these long tail effects in consumption are, they are small compared to other effects of digitization on the products available to consumers. 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 long tail in production rises welfare by over ten times as much as the long tail in consumption. Unpredictability is a generic feature of creative products such as books and movies, suggesting that new products in those industries are generating similar welfare gains. To the extent that product quality is unpredictable in order industries whose entry costs are falling, the effects that we document here may be more widespread.

Our results have some narrow interpretations along with the broader ones. First, the results

[11] We use the following values of s to perform our exercise: 10000000, 5000000, 1000000, 500000, 100000, 10000, 5000, 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 R^2 of 0.39.

in the paper 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. Second, the mechanism articulated in the paper provides a reconciliation of the seemingly contradictory "long tail" and "blockbuster" views of sales concentration since digitization. Reduced costs can foster an ex ante long tail of low expected-value products. But given the unpredictability of product quality at the time of investment, these products' quality realizations can appear throughout the distribution.

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

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.

——— (2010): "Long Tails vs. Superstars: The Effect of Information Technology on Product Variety and Sales Concentration Patterns," *Information Systems Research*, 21, 736–747.

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.

Elberse, A. (2013): *Blockbusters: Hit-making, Risk-taking, and the Big Business of Entertainment*, New York: Henry Holt and Company.

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

IFPI (2013): "Recording Industry in Numbers, The Recorded music Market in 2012," Tech. rep., International Federation of the Phonographic Industry.

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.

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.

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

9 Figures and Tables

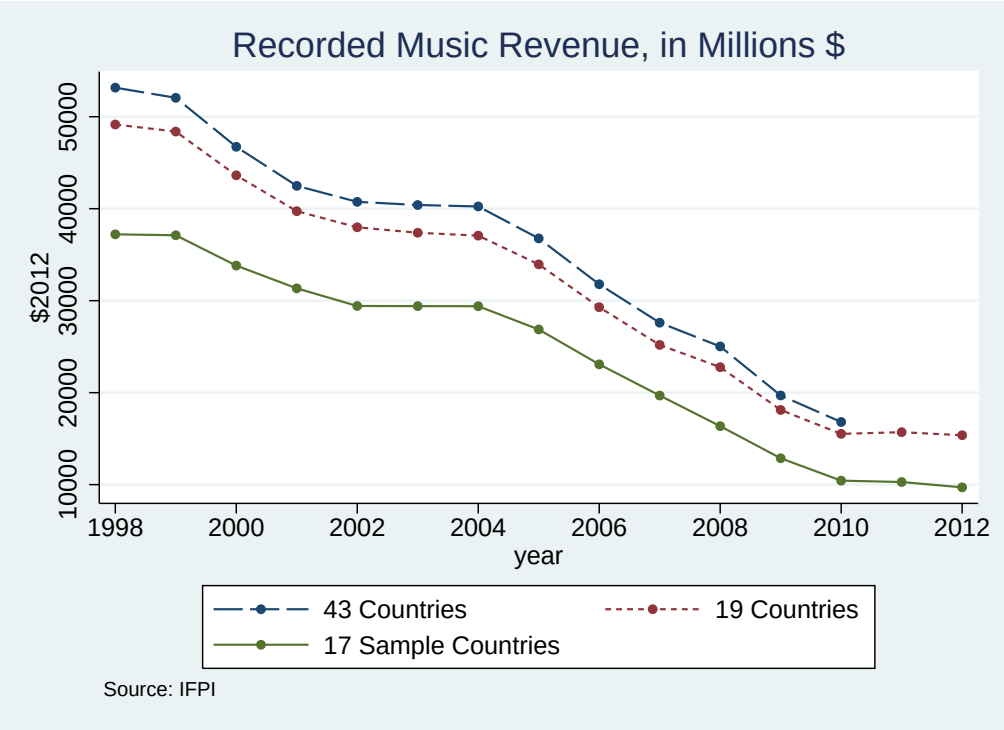

Figure 1: Evolution of Worldwide Music Revenue

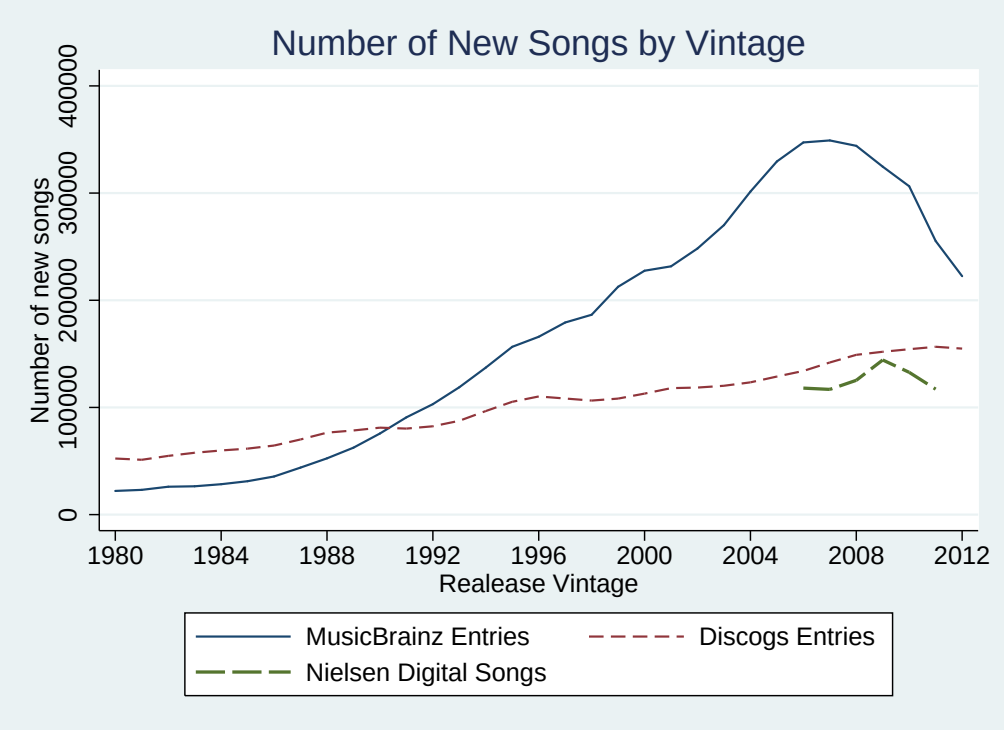

Figure 2: Evolution of the Number of New Songs.

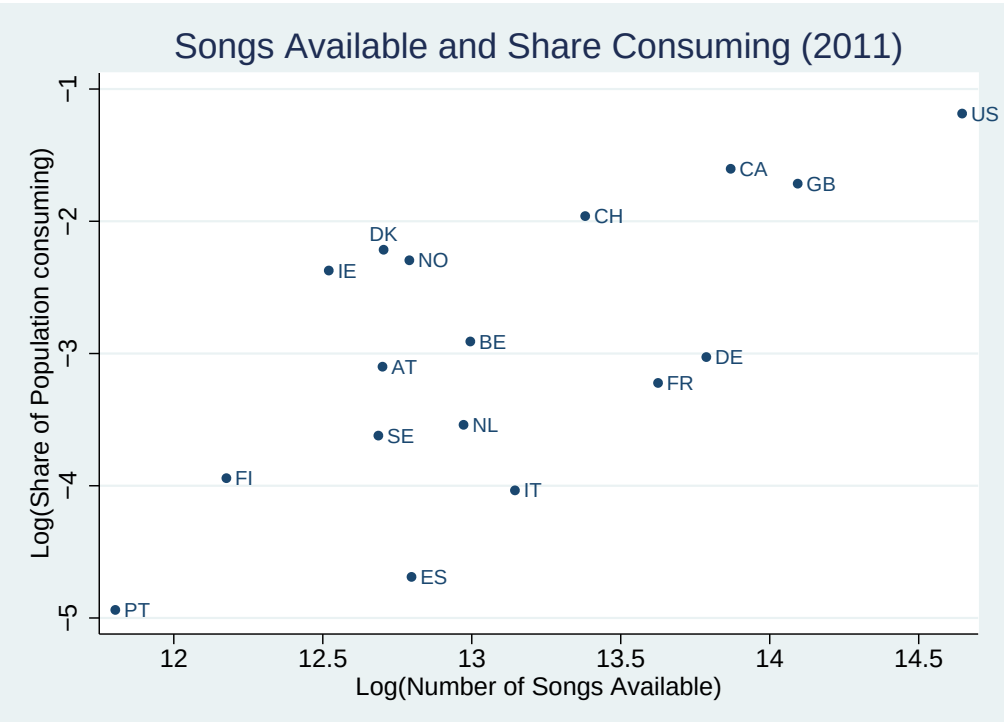

Figure 3: Number of songs available and share of the population consuming, 2011.

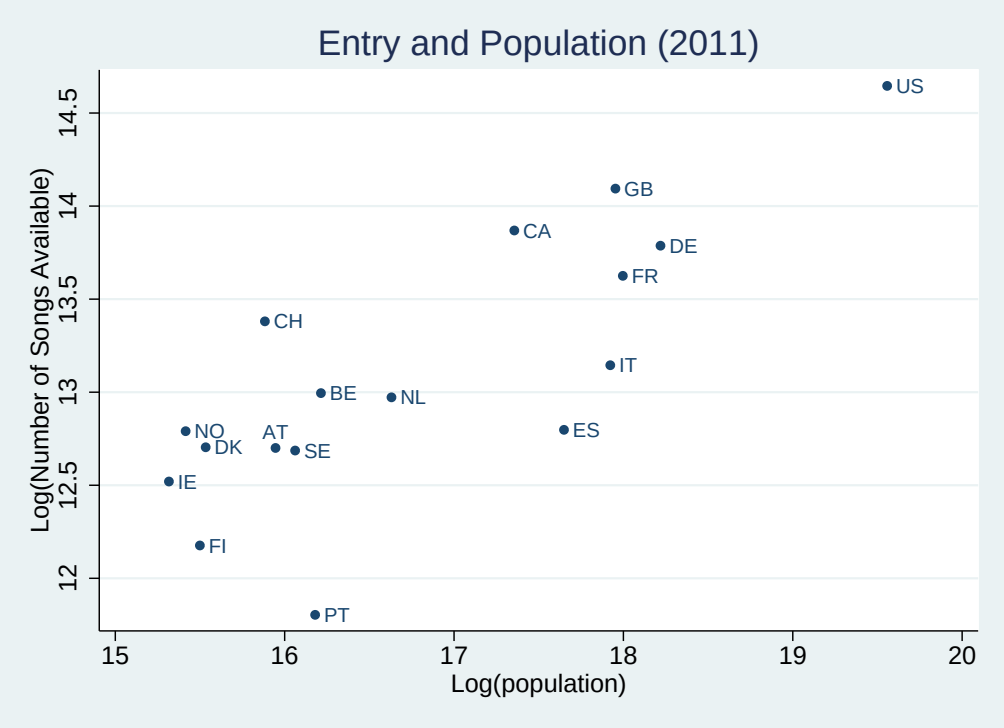

Figure 4: Population and number of songs available, 2011.

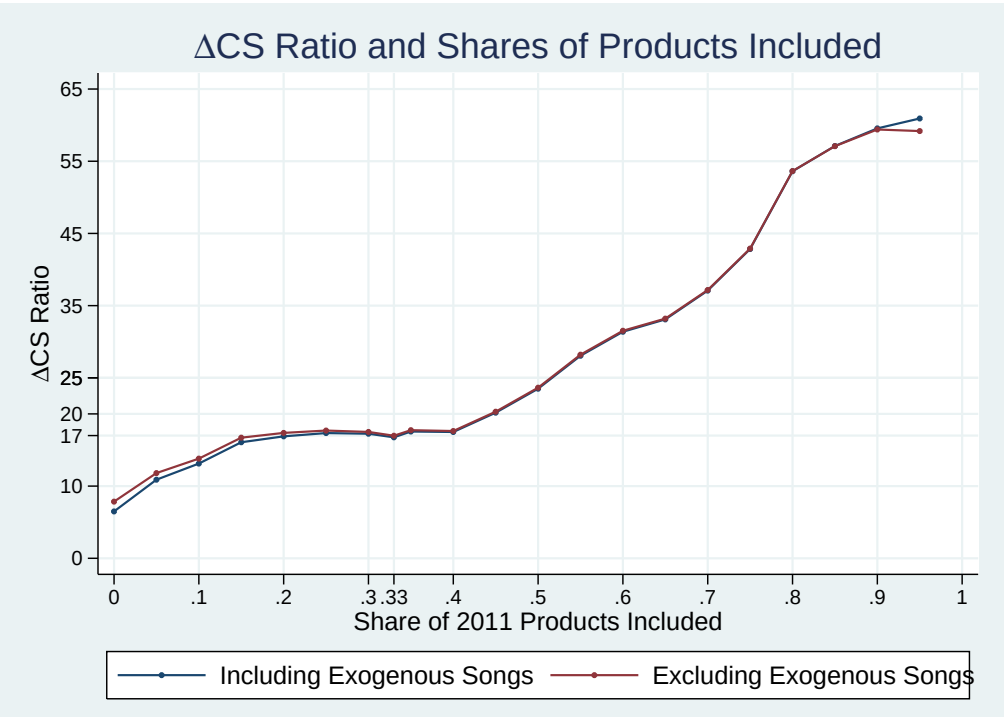

Figure 5: ΔCS Ratio and Shares of Products Included.

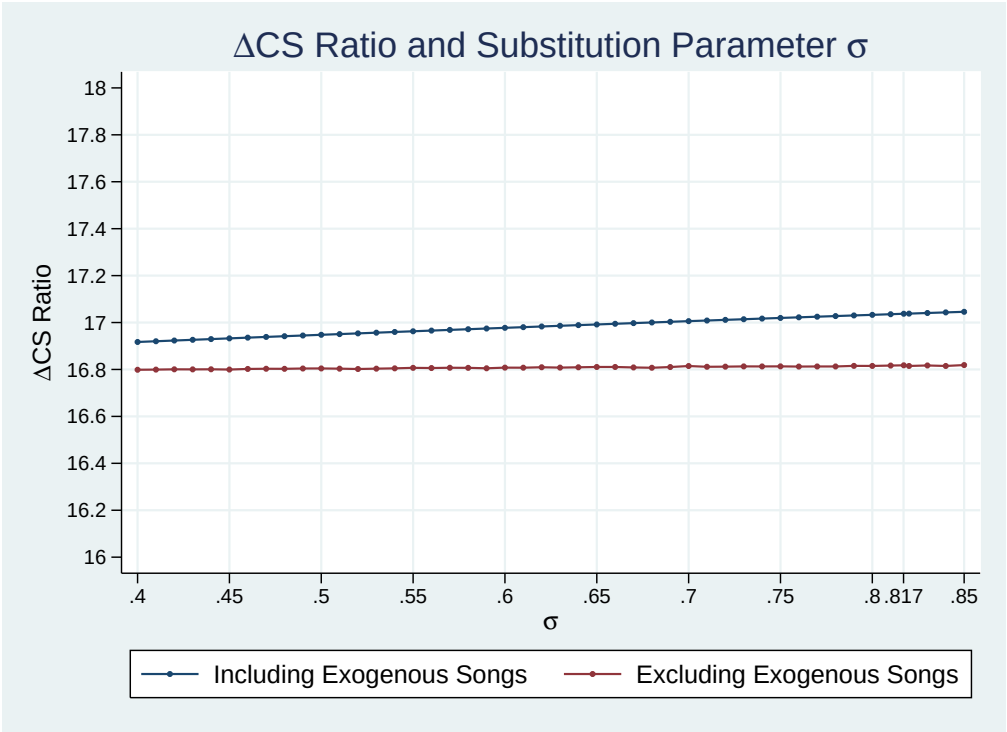

Figure 6: ΔCS Ratio and Substitution Parameter σ.

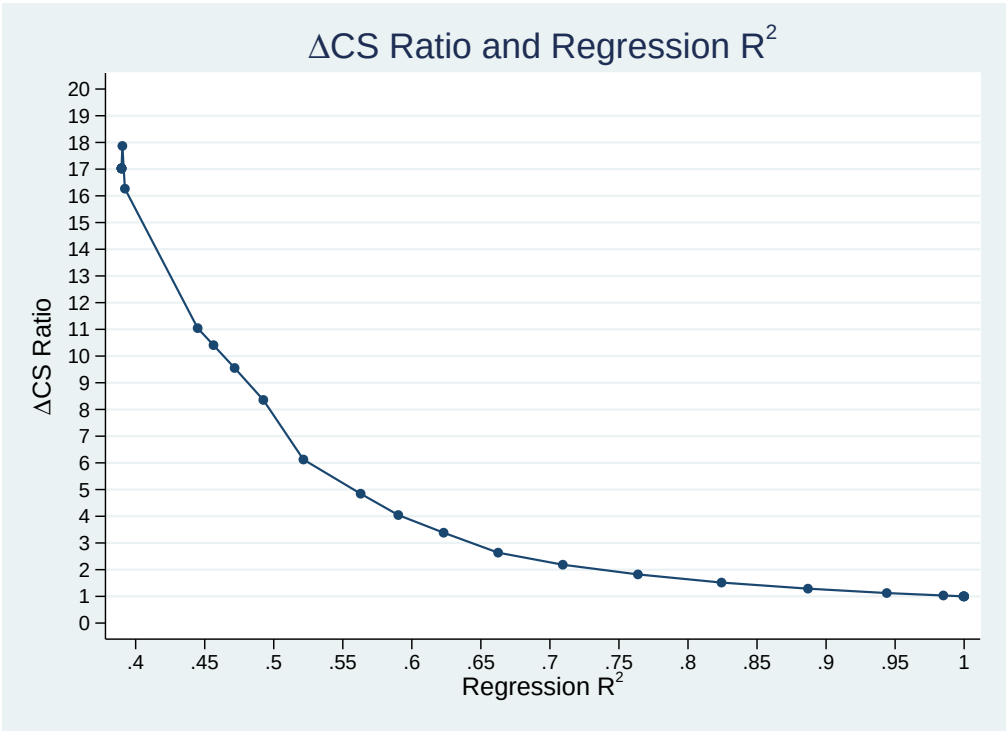

Figure 7: ΔCS Ratio and R^2.

Table 1: Recorded Music Revenue (US$ million, trade value) and Population, 2011.[†]

Country	Population	Physical	Digital	Total (US$)	Digital share	Scaling Factor
Canada	34,482,779	232.5	161.3	393.8	40.96%	2.44
USA	311,591,917	1,841.7	2,344.7	4186.3	56.01%	1.79
Austria	8,419,000	77.0	19.5	96.5	20.23%	4.94
Belgium	11,008,000	97.2	16.5	113.7	14.55%	6.87
Denmark	5,574,000	48.8	30.3	79.1	38.28%	2.61
Finland	5,387,000	44.7	10.8	55.5	19.47%	5.14
France	65,436,552	653.1	186.7	839.7	22.23%	4.50
Germany	81,726,000	1,057.1	208.1	1265.1	16.45%	6.08
Ireland	4,487,000	32.2	16.2	48.4	33.52%	2.98
Italy	60,770,000	152.0	44.1	196.1	22.47%	4.45
Netherlands	16,696,000	158.2	35.5	193.6	18.32%	5.46
Norway	4,952,000	47.0	49.8	96.8	51.41%	1.95
Poland	38,216,000	64.0	4.0	68.0	5.81%	17.21
Portugal	10,637,000	25.0	4.7	29.8	15.92%	6.28
Spain	46,235,000	97.8	42.6	140.4	30.35%	3.30
Sweden	9,453,000	66.4	65.9	132.2	49.80%	2.01
Switzerland	7,907,000	108.3	33.4	141.7	23.59%	4.24
UK	62,641,000	815.5	447.8	1263.3	35.45%	2.82

[†] Source: IFPI, Recording Industry in Numbers (2013).

Table 2: Evolution of the number of sales and number of tracks

	Total Sales (in millions units)						Number of tracks (in thousands)					
Country	2006	2007	2008	2009	2010	2011	2006	2007	2008	2009	2010	2011
Austria	1.08	1.78	2.55	3.42	3.90	4.55	144.60	205.31	261.49	311.39	320.61	327.88
Belgium	2.85	4.63	4.65	5.90	6.57	7.20	223.22	309.88	338.81	404.47	434.31	440.25
Canada	13.94	23.93	37.50	53.57	61.77	83.29	372.86	558.37	728.07	897.65	975.31	1054.95
Denmark	1.94	4.12	5.31	6.39	6.61	7.30	156.25	249.88	283.85	328.94	329.15	329.10
Finland	0.40	0.72	0.89	0.98	0.93	1.25	81.40	122.82	148.49	168.04	158.91	194.16
France	5.21	8.96	18.87	28.32	30.79	31.29	294.51	417.33	600.88	754.79	802.43	826.73
Germany	9.90	23.29	32.92	36.34	41.02	47.53	413.87	661.05	808.85	887.28	920.92	972.09
Ireland	1.28	3.05	3.99	4.53	4.67	5.02	121.50	202.51	241.66	266.54	268.66	273.78
Italy	2.60	4.42	5.74	9.78	10.73	12.90	199.52	284.30	338.93	469.91	495.53	511.39
Netherlands	2.17	2.49	2.81	3.78	4.50	5.82	208.34	279.17	302.52	370.61	402.12	430.36
Norway	1.12	2.85	4.15	5.07	5.49	5.99	146.95	235.16	293.83	344.95	346.82	358.79
Portugal	0.20	0.32	0.65	0.68	0.82	0.91	56.36	81.85	108.43	122.53	130.07	133.74
Spain	1.58	6.23	6.13	5.40	4.85	5.10	149.21	245.26	304.30	341.94	348.50	361.41
Sweden	1.57	2.63	2.96	3.67	3.63	3.04	178.31	261.01	294.35	349.58	349.76	323.36
Switzerland	2.82	4.42	5.79	8.47	10.24	13.36	260.34	352.16	425.09	540.75	573.88	647.41
U.K.	38.21	65.98	96.46	124.07	128.37	135.20	558.11	853.45	1043.27	1223.68	1297.18	1321.07
U.S.	541.44	786.31	994.21	1112.26	1083.70	1142.66	1130.37	1477.00	1787.77	2079.96	2249.62	2294.01

Table 3: Demand Model Estimates (Nested Logit)[†]

	(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(inside share)	0.684***	0.680***	0.781***	0.675***	0.672***	0.817***
	(0.06)	(0.07)	(0.07)	(0.06)	(0.07)	(0.11)
GDP per capita	0.029***	0.029***	0.059***	0.020**	0.020***	0.044**
	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.02)
Share of Digital Sales	4.385***	4.344***	13.545***	3.789***	3.755***	12.635***
	(0.72)	(0.73)	(2.51)	(0.98)	(0.98)	(3.89)
Urban population				-0.002	-0.002	0.015
				(0.01)	(0.01)	(0.02)
Fixed broadband Internet subscribers				-0.018	-0.018	-0.026
				(0.02)	(0.02)	(0.05)
Mobile cellular subscriptions				-0.005*	-0.005*	-0.005
				(0.00)	(0.00)	(0.01)
Internet users				0.031***	0.031***	0.024
				(0.01)	(0.01)	(0.02)
Instrument-year interactions	✗	✓	✓	✗	✓	✓
Covariates-year interactions	✗	✗	✓	✗	✗	✓
Adjusted-R^2	0.829	0.828	0.870	0.844	0.843	0.896
No. of Obs.	50870037	50870037	50870037	50870037	50870037	50870037

[†] Standard errors in parenthesis and clustered at the country level. Inside share instrumented with log(pop) in columns (1) and (3) and with log(pop) interacted with year dummies in specifications (2), (4), (5) and (6). All specifications include year dummies.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table 4: σ Estimates (Nested Logit)[†]

Instrument	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.	(5) Coef./s.e.	(6) Coef./s.e.
Population	0.644***	0.605***	0.876***	0.602***	0.582***	0.815***
	(0.12)	(0.11)	(0.17)	(0.09)	(0.09)	(0.26)
Log(Population)	0.684***	0.680***	0.781***	0.675***	0.672***	0.817***
	(0.06)	(0.07)	(0.07)	(0.06)	(0.07)	(0.11)
Number of Songs	0.417***	0.396***	0.457***	0.434***	0.425***	0.461***
	(0.07)	(0.06)	(0.06)	(0.07)	(0.06)	(0.11)
Log(Number of Songs)	0.412***	0.411***	0.450***	0.446***	0.446***	0.466***
	(0.05)	(0.05)	(0.06)	(0.05)	(0.05)	(0.08)
Full set of covariates	✗	✗	✗	✓	✓	✓
Instrument-year interactions	✗	✓	✓	✗	✓	✓
Covariates-year interactions	✗	✗	✓	✗	✗	✓

[†] Each figure in the table comes from a different estimation of equation (7) and gives the corresponding estimate of σ. Standard errors in parenthesis and clustered at the country level.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table 5: Forecasting Model for New Songs in 2011[†]

	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.
Artist's Age	-12.971***	-10.677***	-20.674***
	(0.26)	(0.24)	(0.64)
Artist's Age squared	0.143***	0.123***	0.310***
	(0.00)	(0.00)	(0.01)
Log(sales in 2009)	70.676***	52.129***	54.058***
	(0.61)	(0.57)	(0.92)
Log(sales in 2008)	-13.224***	-9.568***	-13.411***
	(0.72)	(0.65)	(1.06)
Log(sales in 2007)	-1.579**	-2.661***	-6.012***
	(0.70)	(0.63)	(1.23)
Log(sales in 2006)	5.574***	1.327***	-2.733**
	(0.53)	(0.48)	(1.07)
Years Since Last Release	10.460***	6.703***	18.480***
	(0.45)	(0.41)	(1.48)
New Artist	484.550***	351.116***	317.276***
	(4.26)	(3.94)	(5.48)
Label fixed effects	✗	✓	✓
Age interactions	✗	✗	✓
Age^2 interactions	✗	✗	✓
Years since last release interactions[‡]	✗	✗	✓
R^2	0.196	0.386	0.390
Prediction R^2	0.200	0.315	0.316
No. of Obs.	156411	156411	156411

[†] All estimations use 2010 data and songs from vintage 2010. Interactions are made with both total past sales variables and years since last release. Forecasting R^2 is computed as explained in the text. Standard errors are in parenthesis.

[‡] The years since last release variable is interacted with past sales, artist age and artist age squared.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table 6: Forecasting Models for S Years Old Songs in 2011[†]

	(1) Coef./s.e.	(2) Coef./s.e.	(3) Coef./s.e.	(4) Coef./s.e.
Artist's Age	-20.674***	-29.105***	-15.334***	-11.534***
	(0.64)	(0.49)	(0.52)	(0.54)
Artist's Age squared	0.310***	0.457***	0.271***	0.212***
	(0.01)	(0.01)	(0.01)	(0.01)
Log(sales in 2009)	54.058***	58.628***	90.867***	72.331***
	(0.92)	(0.57)	(0.87)	(0.56)
Log(sales in 2008)	-13.411***	-12.963***	-20.379***	
	(1.06)	(0.74)	(0.69)	
Log(sales in 2007)	-6.012***	-7.864***		
	(1.23)	(0.69)		
Log(sales in 2006)	-2.733**			
	(1.07)			
Years Since Last Release	18.480***	-12.521***	-12.110***	-22.598***
	(1.48)	(1.11)	(1.04)	(1.15)
New Artist	317.276***			
	(5.48)			
Label fixed effects	✓	✓	✓	✓
Age interactions	✓	✓	✓	✓
Age^2 interactions	✓	✓	✓	✓
Years since last release interactions[‡]	✓	✓	✓	✓
R^2	0.390	0.416	0.447	0.456
Prediction R^2	0.316	0.340	0.381	0.359
No. of Obs.	156411	229330	213393	209444

[†] All estimations use 2010 data. Specification (S) is used to predict the quality of (S-1) year old music in 2011. Interactions are made with both total past sales variables and years since last release. Standard errors in parenthesis.

[‡] The years since last release variable is interacted with past sales, artist age and artist age squared.

* Significant at the 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

Table 7: Fixed Costs of Entry[†]

	Vintages 2008-2011			Vintage 2011		
Regime	PF	Imp Pred	No Pred	PF	Imp Pred	No Pred
Counterfactual	17.91	224.80	1280.66	22.01	203.85	1740.05
Status Quo	1.00	15.47	858.13	1.00	44.64	1555.28

[†] PF: Perfect Foresight; Imp Pred: Imperfect Predictability; No Pred: No Predictability.

Table 8: Counterfactual Results[†]

Songs	Regime	ΔCS	Ratio CS	ΔRev	Ratio Rev	ΔTC.	Ratio TC	ΔW	Ratio W
Vint 2008-2011	PF	0.30	1	0.30	1	-3.66	1	4.26	1
Vint 2008-2011	Imp Pred	3.33	11.16	3.33	11.17	-43.81	11.96	50.47	11.85
Vint 2008-2011	No Pred	65.48	219.71	66.32	222.51	317.73	-86.75	-185.93	-43.66
Vint 2011	PF	0.06	1	0.06	1	-0.85	1	0.97	1
Vint 2011	Imp Pred	1.03	16.82	1.03	16.83	-3.13	3.68	5.20	5.34
Vint 2011	No Pred	18.62	302.93	18.68	304.17	130.92	-153.92	-93.62	-96.18

[†] All figures are in million of $. The columns labeled Ratio report the ratio between Δ in the corresponding regime and Δ in the Perfect Foresight Regime. PF: Perfect Foresight; Imp Pred: Imperfect Predictability; No Pred: No Predictability.