

Competition Between Local and Electronic Markets: How the Benefit of Buying Online Depends on Where You Live

Chris Forman

College of Management, Georgia Institute of Technology, Atlanta, Georgia 30308,
chris.forman@mgt.gatech.edu

Anindya Ghose

Stern School of Business, New York University, New York, New York 10012, aghose@stern.nyu.edu

Avi Goldfarb

Rotman School of Management, University of Toronto, Toronto, Ontario M5S 3E6, Canada,
agoldfarb@rotman.utoronto.ca

Our paper shows that the parameters in existing theoretical models of channel substitution such as offline transportation cost, online disutility cost, and the prices of online and offline retailers interact to determine consumer choice of channels. In this way, our results provide empirical support for many such models. In particular, we empirically examine the trade-off between the benefits of buying online and the benefits of buying in a local retail store. How does a consumer's physical location shape the relative benefits of buying from the online world? We explore this problem using data from Amazon.com on the top-selling books for 1,497 unique locations in the United States for 10 months ending in January 2006. We show that when a store opens locally, people substitute away from online purchasing, even controlling for product-specific preferences by location. These estimates are economically large, suggesting that the disutility costs of purchasing online are substantial and that offline transportation costs matter. We also show that offline entry decreases consumers' sensitivity to online price discounts. However, we find no consistent evidence that the breadth of the product line at a local retail store affects purchases.

Key words: channel substitution; theory testing; Internet retailing

History: Accepted by Barrie R. Nault, information systems; received November 30, 2006. This paper was with the authors 6 months for 3 revisions. Published online in *Articles in Advance* November 26, 2008.

1. Introduction

As of 2006, electronic commerce represented just 3% of total retail sales (U.S. Census Bureau 2007). Online shopping remains a small fraction of retail sales despite the well-known benefits of electronic commerce to consumers, including lower prices (e.g., Brynjolfsson and Smith 2000), greater selection and availability (e.g., Ghose et al. 2006), and greater convenience by eliminating travel costs and enabling 24×7 purchases irrespective of geographic location (Cairncross 1997). Of course, there are many reasons why consumers do not buy online: inspecting nondigital products is often difficult, shipping can be slow and expensive, and returning products can be challenging. That is, there appears to be a set of fixed disutility costs of buying online. These costs vary across products and retailers, and in some markets have created significant hurdles to the continued diffusion of electronic commerce.

Theoretical research has explored consumer channel choice in commodity markets, modeling the decision as a trade-off between these fixed disutility costs and the lower search and transportation costs of buying online, in addition to any price differences across the two channels (starting with Balasubramanian 1998). However, there is no systematic empirical evidence on the trade-off between offline transportation costs and online disutility costs. In short, while theory often assumes that the benefits of buying online depend on where you live, we do not know how much this matters. In exploring the online-offline trade-off, our paper is in the spirit of prior research that has provided empirical validation to theories on how the Internet influences buyer decisions due to lower search costs (Brynjolfsson and Smith 2000), greater product selection (Brynjolfsson et al. 2003), and information about word-of-mouth based on user-generated reviews (Chevalier and Mayzlin 2006, Forman et al. 2008).

Using monthly data from Amazon.com on top-selling books in 1,497 local markets over the 10 months from April 2005 to January 2006, we show that distance to a local store matters in online purchasing and that distance mitigates online price effects. In particular, we examine how entry by Wal-Mart, Target, Barnes and Noble, and Borders changes the types of products bought online in the location where the store entered and compare this to the types of products bought in locations that did not experience such entry. Our method controls for differences in consumer preferences across locations through product-location fixed effects. Thus, we use store entry to identify the effects of improved offline options on online choice using a difference-in-difference strategy. By focusing on books, we study a commodity product where brand-specific and product-specific factors are less likely to influence channel substitution, and where purchase-related factors that cannot be determined digitally (Lal and Sarvary 1999) are relatively unimportant. Moreover, e-commerce book sales are high (so the trade-off we explore is economically important) and online disutility costs are relatively low (so our estimates of online disutility costs relative to offline transportation costs are likely to be conservative, in comparison to other products). Books also have the advantage of having one dominant online retailer (Amazon.com) and easily identified offline retailers making it possible to identify the appropriate scope for the study.

We find that people substitute away from online purchasing toward offline purchasing when a store opens locally: people appear to respond to increased convenience in the offline channel. After a discount retailer (Wal-Mart/Target) or a large specialty store (Barnes and Noble/Borders) enters a market, local online purchases of the nationally most popular products decline relative to purchases of products unlikely to be prominent, or even available, offline. These effects are economically large, suggesting substantial disutility costs of purchasing online, even for books. We also show that offline entry decreases consumers' sensitivity to online price discounts. However, we do not find consistent evidence that the breadth of the product line at a local retail store affects purchases. Although Barnes and Noble has a much wider selection of books than Wal-Mart, entry by either has the same primary effect: the most popular products become less likely to be bought online. We attribute this to high offline transportation costs (in expectation) due to uncertain availability of less popular books at offline stores and limited consumer demand for less popular products. However, we do find evidence that offline product selection matters in locations with a university and larger cities, where consumer tastes may be more varied and therefore the

concentration of consumers with preferences for less popular products is likely higher.

Our paper contributes to three areas of research. First, and most importantly, we provide empirical support for assumptions widely used in theoretical models of online-offline channel substitution (Balasubramanian 1998, Pan et al. 2002, Jeffers and Nault 2007, Viswanathan 2005, Chun and Kim 2005, Liu et al. 2006, Moorthy and Zhang 2007, Guo and Liu 2008, Cheng and Nault 2007, and others). By providing evidence for the importance of transportation costs and online disutility costs and shedding light on their relative magnitudes, we provide further insights into results in these papers that often depend on these parameter values.

Second, our paper contributes to a small empirical literature on consumer substitution between online and offline channels (Goolsbee 2001, Ellison and Ellison 2006, Prince 2007). Most of this prior work focuses on cross-price elasticities; our paper explores how offline retail location affects online purchases. Although Brynjolfsson et al. (2008) do examine the role of local characteristics in women's clothing, they focus on how equilibrium market conditions relate to online choices in a cross section. In contrast, the panel nature of our data means that we can separately identify local demand-side preferences from supply-side factors related to retail competition.

Third, and more broadly, this paper advances the emerging empirical literature that studies how online retailing contributes to consumer welfare. Various streams of this literature have shown how Internet retailing benefits consumers with lower prices (surveyed in Baye et al. 2006), lower search costs (Brynjolfsson et al. 2003), higher resale values of new products by providing a more liquid market for used books (Ghose et al. 2005, 2006) and better information about location-specific product preferences through user-generated opinions (Forman et al. 2008). We contribute to this literature by examining the benefits of Internet retailing in improving customer convenience.

2. Hypotheses

Our hypotheses build on existing theoretical models that examine consumer substitution between online and offline channels. In particular, our paper is closely related to research on multichannel retailing that utilizes theoretical models of spatially differentiated commodity markets derived from Salop's (1979) circular city model (Balasubramanian 1998, Jeffers and Nault 2007, Viswanathan 2005, Cheng and Nault 2007, Guo and Liu 2008) and from Hotelling's (1929) linear city model (Pan et al. 2002, Chun and Kim 2005, Liu et al. 2006, Moorthy and Zhang 2007). Common assumptions in all of these models are the presence of

transportation costs when consumers use the offline channel and of disutility costs when buying online. In some cases the size of the transportation costs plays a key role in determining the equilibrium that prevails in these models.

As noted above, the core conceptual framework in our paper is derived from spatial models of competition that include a direct marketer, in particular Balasubramanian's (1998) circular city model of offline retailers with a direct retailer in the center. This model includes several key assumptions that motivate our first hypothesis. Consumers buy a single standard product, and have complete information about prices and product availability. Consumers face a finite cost of traveling to traditional retailers that depends on their distance to the retailer. Therefore, consumers have heterogeneous costs of buying offline that depend on their location. These costs may be monetary costs of travel, inconvenience costs, and/or the opportunity cost of time. Consumers have a high reservation price relative to their transportation cost and the product is known to be in stock (so the market is covered and the product is "popular"). In contrast, all consumers face an identical fixed cost of buying from a direct or online channel (e.g., a shipping cost, an inability to assess product quality, or a lack of immediate gratification). Furthermore, in contrast to Viswanathan (2005), there are no switching costs or network externalities that reduce channel switching.

Consumers maximize utility by choosing between the offline and online retailer based on prices, offline transportation costs, and online disutility costs. All else equal, reductions in transportation costs directly increase the utility of purchasing from the offline retailer, and therefore decrease the likelihood that the representative consumer buys from the online retailer. To our knowledge, this direct test of the role of distance in the Balasubramanian (1998) model has not been performed in any prior work.

HYPOTHESIS 1A (CONVENIENCE FOR POPULAR PRODUCTS). *As distance to offline stores decreases, the likelihood of purchasing a commodity product online decreases.*

We also examine the impact of distance to the offline retailer on products that are not stocked in all offline stores. We label such products "less popular." Although not previously emphasized in the literature, product selection may be an important factor in channel choice. Hypothesis 1A assumes consumers are fully informed about the price and availability of products in both channels. This setting is similar to the market for best-selling books. For less popular products, consumers are less certain about the availability of the product at the offline retailer. This

can be viewed as an increase in average offline transportation costs (in expectation) for a given product. As Cheng and Nault (2007) note, an example of such a market might be that for ethnic books in the United States. In such a setting, reduction in the distance to offline stores has a weaker effect on the likelihood a representative consumer buys online for two reasons. First, the reservation value of the representative consumer is lower, so changes in transportation costs have a smaller impact on the likelihood of buying online. Second, the likelihood that any given store has the less popular product is smaller, so the expected transportation cost declines less than if the product was a popular one (and was certain to be available at the offline retailer).

HYPOTHESIS 1B (PRODUCT SELECTION). *As distance to offline stores decreases, online purchases of a commodity product that is highly likely to be stocked offline decrease more than online purchases of a commodity product that is less likely to be stocked offline.*

For example, take a book that is likely sold at a large specialty store such as Barnes and Noble but not at a discount store such as Wal-Mart. Hypothesis 1B implies that the effect of Barnes and Noble on online sales of this book is larger than the effect of Wal-Mart. It is a version of the convenience Hypothesis 1A, but it takes into account the fact that not all kinds of stores stock all products.

Our final hypothesis examines the role of online and offline prices on channel choice. In Balasubramanian's (1998) model, changes in online price directly influence the utility of buying offline, and vice-versa. That is, there exists a significant cross-price elasticity across the online and offline channels. Prior work has tested for and found such a cross-price elasticity in computers (Prince 2007) and computer memory (Ellison and Ellison 2006), so we incorporate cross-price elasticity in our econometric model but do not include it as a separate hypothesis. Instead, we focus on how distance to retail stores is associated with changes in consumers' sensitivity to price. Decreases in distance to offline stores will, as before, increase the utility of buying offline. This makes a given representative consumer less sensitive to changes in online price. So, a marginal consumer who would have previously switched to the online channel after a fall in the online price no longer does so. Therefore, the impact of online discounts is tempered by the existence of local retail stores.

HYPOTHESIS 2 (PRICE). *As distance to offline stores decreases, online price decreases have a smaller (less positive in magnitude) impact on the likelihood of purchasing a commodity product online relative to the change from a price decrease made prior to the decrease in distance.*

3. Data Description

To examine how online behavior varies with offline supply conditions, we require detailed data on how consumer online purchases vary across local geographic markets. The data we use are online book purchases from Amazon.com. Books are a particularly good setting to test our hypotheses for several reasons. First, books are commodity products wherein brand-specific or product-specific factors are less likely to influence consumer substitution across channels. Second, purchase-related attributes that cannot be determined digitally (Lal and Sarvary 1999) are relatively unimportant in the book market, enabling us to focus on location-related factors. Third, because books are inexpensive commodity products, they are representative of a wide variety of other commodity products available online, including DVDs, CDs, groceries, office products, and others. Fourth, books are one of the few product categories (besides travel services and computer hardware) where online sales reached over 10% of total retail sales by 2005 (U.S. Census Bureau 2007). And finally, the main offline book retailers are easy to identify, and we have precise data on when these stores open in a given location. Consequently, we can set up an effective natural experiment to explore channel substitution.

3.1. Raw Data from Amazon

An observation in our data consists of a particular product-location-month. The raw data come from the webpages on “Purchase Circles” from the Amazon.com website. Amazon’s Purchase Circles are specialized best-seller lists that denote the top-selling books by location throughout the United States. Henceforth, we use the word *locations* to refer to small and large cities, as well as small towns. When deciding upon the length of our sample, it was important that our time series be able to separate the short-run (for example, due to curiosity effects on the part of households) from the long-run effect of entry. Singh et al. (2006) examine the effects of Wal-Mart entry on local (offline) supermarket sales and compare the short-run effects to the long-run (defined as longer than three months) effects. They find the long-run effect on store visits is slightly larger than that on short-run visits, but overall effects on expenditures are slightly greater in the short run (–18.5% in the short run versus –17.8% in the long run). We collect data between April 2005 and January 2006, a 10 month period that allows us to separate the short-run from the longer-run effects that persist after three months. We used a JAVA “spider” to visit Amazon’s website and collect monthly data on purchases for each location in the Purchase Circles.¹

¹ Some locations in our Purchase Circles raw data set do not appear for the entire time period. In particular, due to a managerial

To be included as a Purchase Circles location, the number of purchases in a location needs to be above a threshold. Therefore, the use of Purchase Circles means that we do not have a census (or a truly random sample) of locations in the United States. To better understand the consequences of using this data, we matched the locations in our data to 2,000 U.S. Census place data using place names. The smallest location in our data is Weldon Spring Heights, Missouri (place population 79). The largest location in our data is Los Angeles, California, with a place population of 3,694,820 (both Amazon and the U.S. Census Bureau divide New York City into neighborhoods). Our data constitute 50.3% of the total place population and 60.8% of the place population exceeding 10,000. Among places with population greater than 10,000, median household income in our data is \$52,268 compared to \$47,107 nationwide; population is 99,605 compared to 50,830. Thus, while our data does tend to oversample locations with higher than average population and income, we do have information on smaller locations (269 of the locations in our sample have under 10,000 people). Despite these limitations, to our knowledge these data provide the most representative source of cross-sectional online purchase behavior available.

Next, we describe the construction of our variables. Further details are provided in the online appendix (available in the e-companion).² Descriptive statistics are provided in Table 1.

3.2. Dependent Variable

For each location, Amazon provides a list of the top 10 selling products. Our primary dependent variable, $LocalTop10_{ijt}$, is a binary variable that is equal to one if book i is present in the local top 10 in location j in month t , and zero otherwise. Though our data contain only information on the products that appear in the top 10 in a location, there is considerable heterogeneity in this measure across locations and over time. Consumers buy different products in different locations; 58.6% of products in our sample appear in the top 10 products at five or fewer locations.

The use of rank data, rather than quantity data, means that our empirical framework is different from those typically used to examine channel substitution: Our analysis is based on relative rather than absolute sales. Therefore, we translate our hypotheses into testable implications of how the relative sales of

decision at Amazon related to the threshold for inclusion in Purchase Circles, the number of locations expanded in November 2005. For this reason, we only include locations that are observed before and after this date. This resulted in 1,497 locations.

² An electronic companion to this paper is available as part of the online version that can be found at <http://mansci.journal.informs.org/>.

Table 1 Summary Statistics for Books

Variable	Observations	Mean	Standard deviation	Minimum	Maximum
By location-product-month					
Dummy for top 10 in location	4,051,254	0.0347	0.1831	0	1
Relative price	4,051,254	-0.2654	0.1434	-0.6	0
Very popular products (rank 1–150)	4,051,254	0.1711	0.3766	0	1
Popular products (rank 150–500)	4,051,254	0.1737	0.3789	0	1
Moderately popular products (rank 500–1,500)	4,051,254	0.1538	0.3608	0	1
Somewhat less popular products (rank 1,500–5,000)	4,051,254	0.1351	0.3418	0	1
Less popular products (rank 5,000–15,000)	4,051,254	0.1296	0.3358	0	1
Unpopular products (rank over 15,000)	4,051,254	0.2367	0.4251	0	1
Dummy for missing price information	4,051,254	0.0644	0.2454	0	1
Average rating	4,051,254	4.1098	0.5617	1.5	5
Log(days since launch)	4,051,254	6.5007	1.4946	0	9.8268
Broadband	4,051,254	11.4887	3.3362	0	24
Dummy for missing elapsed date information	4,051,254	0.0259	0.1588	0	1
Log(Number of reviews)	4,051,254	4.9545	1.4596	0.6931	8.6500
Discount store entry within 5.4 miles	4,051,254	0.0809	0.2727	0	1
Large bookstore entry within 5.4 miles	4,051,254	0.0166	0.1276	0	1
By location					
Discount store openings in all locations	1,497	0.1643	0.3707	0	1
Discount store openings in small locations	143	0.0979	0.2982	0	1
Discount store openings in large locations	412	0.2087	0.4069	0	1
Large bookstore openings in all locations	1,497	0.0468	0.2112	0	1
Large bookstore openings in small locations	143	0.0210	0.1438	0	1
Large bookstore openings in large locations	412	0.0752	0.2641	0	1
Location has a university	1,497	0.4449	0.4971	0	1

Notes. Unit of observation in the top half of the table is a location-product-month. Unit of observation in the bottom half of the table is a location.

popular and less popular products vary across locations. These testable implications arise from the fact that while sales of popular products are sensitive to variations in local retail store distance, sales of unpopular products that are not stocked in local retail stores are not. Hypotheses on the likelihood that a particular book is purchased online are therefore translated into testable implications of the likelihood that a particular product appears in a local top 10.

3.3. Product Characteristics

We use information on product details from Amazon’s website. For each book listed in Purchase Circles, we collected data from Amazon on the list price, Amazon’s retail price, the product’s national sales rank on Amazon, the product’s release date, the average rating from Amazon’s customers, and the number of reviews posted on Amazon. Shipping costs are identical across locations and are therefore not collected.

To measure the price benefits of online retailing, we construct another variable that we label *RelativePrice*. The *RelativePrice* variable is computed as the difference between the Amazon retail price and the undiscounted list price, normalized by the list price. The list price is the “recommended” price for a book that is typically printed on the book itself. Bestsellers are usually discounted, both online and offline, while older products typically sell at the list price.

In addition to price, we examine the national rank (popularity) of a book on Amazon. To allow for a flexible functional form, we compute a series of dummy variables (a spline) that indicate the specific range of national sales rank for which the book appears in that month: top 150, 151–500, 501–1,500, 1,501–5,000, 5,001–15,000, or greater than 15,000 (which we use as the base). We define very popular products as those that fall in the top 150 nationally and popular products as those that fall in the range 151–500. Products with national sales ranks in the lower ranges, specifically those not in the top 1,500, are classified as somewhat less popular (1,501–5,000) and less popular (5,001–15,000) products. Although our results are robust to a log-linear specification and to using *New York Times* and *USA Today* bestsellers lists, we focus on the Amazon rankings because they provide detail on the rank of all products and allow for differences between popular and less popular products.

To construct our final data set, we identified the 300 products that were most frequently listed in the local top 10 lists in each month. We added an “outside option” of products listed in a local top 10 but not in this group of 300. This outside option had characteristics equal to the average of its products. The choice of 300 was based on a trade-off between two competing objectives. To identify whether product selection matters (Hypothesis 1B), we wanted to make the size of the choice set as large as possible;

however, if we made the choice set too large, then we would have many products that are rarely in a local top 10 which is unappealing due to the product-location fixed effects. We judged 300 the best compromise in this trade-off, though our results are robust to other specifications.

3.4. Store Entry and Location-Level Data

Our main analysis examines how offline retail store entry influences buyer choice online. Retail store entry in a given location decreases the average distance consumers in that location must travel to access offline retailers, other things equal. We examine entry of two types of stores. For each location in our data set, the variable labeled *DiscountStoreEntry* is equal to one for every month after a Wal-Mart or Target store has entered within a 5.4 mile radius of the location and zero otherwise; our variable labeled *Large-BookstoreEntry* is equal to one for every month after a Barnes and Noble or Borders bookstore has entered within a 5.4 mile radius of the location and zero otherwise. These data were collected through press releases from the companies and through direct communication with company representatives. To compute radii, we use the average longitude and latitude across zip codes within the location. We use 5.4 miles because this is the distance that the average consumer travels to go to a bookstore (Brynjolfsson and Smith 2000), although the results are robust (and in fact stronger) when we use a larger radius of 20 miles. Across our entire sample, 16.4% of locations experience discount store entry, whereas 4.7% experience a large bookstore entry. We focus on these particular stores because they represent the top two bookstores and the top two retailers who sell books, with a substantial drop-off in sales for the third-place retailer.

In addition to the store entry data, we also collected information on population using the U.S. Census Bureau estimates for 2004, on whether each location has a university from Barron's educational series, and on the number of broadband providers in each location from semiannual Federal Communications Commission Form 477 data from December 2004, June 2005, and December 2005.

4. Econometric Model

As discussed above, we examine the trade-off between the transportation and search costs of buying offline and the various disutility costs of buying online. Identifying this trade-off, however, is challenging because it is difficult to separately identify supply and demand effects. For example, large cities may differ from small towns because there are more stores in large cities (supply) or people in large cities and small towns have different tastes (demand). One solution is to directly measure the number of stores in each

location and to regress sales rank on number of stores and include demographics to attempt to control for taste. However, this would likely suffer from the same difficulty: locations with more bookstores are likely those locations where many people buy books; there are more bookstores because of local tastes. Thus, separating out the effect of interest (how local competition affects online purchases) from other effects such as demand variation cannot be done in a simple cross-section.

A common solution in the economics literature is to use instrumental variables: if we could identify something that is correlated with the number of stores in a market but not with local demand then we could use that to identify the effect of the number of stores on online purchases. Unfortunately, we do not have access to such an instrument. Local characteristics associated with the number of stores selling books (e.g., population or education) are likely correlated with local preferences for books.

Therefore, we use an alternative technique for causal inference: difference-in-differences. The basic idea of difference-in-differences is to examine a set of treated units before and after the treatment (in this case, store entry). Given that many other factors may have changed around this time, we use a control group (places with no store entry) to control for these factors and isolate (to the best of our ability) the effect of the treatment. The regression approach to difference-in-differences also allows for regression controls. Thus, indexing units by j and time by t , we adopt the basic framework:

$$\begin{aligned} Outcome_{jt} = & \beta_0 + \beta_1 TreatmentGroup_j + \beta_2 AfterTreatment_t \\ & + \beta_3 TreatmentGroup_j * AfterTreatment_t \\ & + \gamma RegressionControls_{jt} + \varepsilon_{jt}. \end{aligned} \quad (1)$$

By plugging in zeros and ones for the binary variables in Equation (1), the difference across groups in the before-after treatment is clearly β_3 . If β_3 is positive, the treatment can be interpreted as having a positive effect on the outcome. Just as in a true experiment, this "natural experiment" approach means we see whether behavior in the treatment group changes differently from behavior in the control.

In our case, $Outcome_{it}$ is whether a book is in the online local top 10 and the treatment is whether a store entered. $TreatmentGroup_j$ is the set of locations that experience store entry. $AfterTreatment_t$ measures whether the store has entered by time t . Under some identifying assumptions (described below), this method allows us to establish how much store entry attracts consumers away from Amazon and toward

the offline channel. This gives us our estimating equation:

$$\begin{aligned}
 & (\text{LocalTop10}_{ijt}) \\
 & = \alpha_0 + \alpha_1 \text{DiscountStoreEntry}_{jt} + \alpha_2 \text{LargeStoreEntry}_{jt} \\
 & \quad + \beta \text{NationalRank}_{it} + \gamma \text{NationalRank}_{it} \\
 & \quad \times \text{DiscountStoreEntry}_{jt} + \delta \text{NationalRank}_{it} \\
 & \quad \times \text{LargeStoreEntry}_{jt} + \theta_1 \text{RelativePrice}_{it} \\
 & \quad + \theta_2 \text{RelativePrice}_{it} \times \text{DiscountStoreEntry}_{jt} \\
 & \quad + \theta_3 \text{RelativePrice}_{it} \times \text{LargeStoreEntry}_{jt} \\
 & \quad + \phi X_{it} + \mu_{ij} + \mu_t + \varepsilon_{ijt}. \tag{2}
 \end{aligned}$$

Here (LocalTop10_{ijt}) is a dummy variable for whether product i is in the top 10 in location j for month t ; $\text{DiscountStoreEntry}_{jt}$ and $\text{LargeStoreEntry}_{jt}$ indicate whether a discount store or large bookstore entered location j in month t or earlier; NationalRank_{it} is a vector of dummy variables for the national sales rank of product i in month t ; $\text{RelativePrice}_{it}$ is the online price relative to the list price; X_{it} are other attributes;³ μ_{ij} is a product-location fixed effect; μ_t is a month fixed effect; and ε_{ijt} is a product-location-month idiosyncratic error term. The product-location fixed effect, μ_{ij} , controls for all time-invariant location-specific preferences and is key to the difference-in-difference identification.

The key assumption in difference-in-difference estimation is that unmeasured factors affect the treatment and control groups equally. Although the product-location fixed effects in our model control for possible differences between the treatment locations (that experience entry) and the control locations (that do not), if areas that experience entry also experience a change in local demand preferences then the treatment group changes over time differently than the control group. We believe this assumption is reasonable given our rich econometric controls and relatively short time period.

There are two additional properties of our empirical framework that are important to discuss here. First, our coefficients of interest are on interaction terms. This means that nonlinear models (e.g., Probit) are difficult to interpret because the cross-partial may have a different sign than the coefficient on the interaction term (Ai and Norton 2003). The main disadvantage of using a linear model is reduced efficiency. Given the large number of observations in our

study, this is less important. Second, our difference-in-difference estimates may overstate the significance of the results without a standard error correction that addresses the fact that a given location is counted several times (i.e., for many products) even though entry occurs just once (Bertrand et al. 2004). For this reason, we cluster by location-month and use heteroskedasticity-robust standard errors.⁴

Our hypotheses from §2 easily convert into testable hypotheses on the coefficients of the interactions. Table 2 summarizes these coefficients and our results. Hypothesis 1A suggests that decreases in distance to offline stores are associated with fewer purchases of popular products online. Entry by any type of store decreases such distances, other things equal. Therefore the coefficients on the interactions of $\text{DiscountStoreEntry}$ or LargeStoreEntry with NationalRank dummies for products that are nationally in the top 150 and in the 151–500 range are hypothesized to be negative. Hypothesis 1B looks at product selection. Since large bookstores have a larger selection than discount stores, we expect large bookstore entry to have a larger impact on the less popular (i.e., nationally ranked in the 5,000–15,000 range) and somewhat less popular products (in the 1,500–5,000 range) than discount store entry. We chose this range because the typical Wal-Mart has under 2,000 books and the typical specialty bookstore has a much higher number.⁵ So, we expect the coefficient on LargeStoreEntry interacted with NationalRank products in the 5,000–15,000 range to be more negative than the coefficient on $\text{DiscountStoreEntry}$ interacted with the products in this range. Hypothesis 2 suggests that store entry mitigates the effect of online price discounts because offline retailers discount the same types of books as the online retailer: i.e., the interactions of $\text{DiscountStoreEntry}$ or LargeStoreEntry with RelativePrice will be positive.

5. Results

In this section, we show that changes in distance to local retail stores have a substantial effect on the types of products that appear in a local top 10 list. Our main results are in Table 3, column 1. Column 2 shows robustness to an alternative measure of distance; many further robustness checks are available in

⁴ Robust standard errors also address the possibility that the error differs by location size because local popularity rankings could have different random components in smaller locations. This would lead to measurement error in the dependent variable, thereby adding heteroskedasticity to the error term.

⁵ The average Barnes and Noble stocks between 60,000 and 200,000 books. Approximately 50,000 of these books are common across all stores (Rosenthal 2005). For discount stores, Wagner (2003) listed the number to be between 1,000 and 1,500, a number corroborated by our own personal survey of stores in Atlanta and New Jersey.

³ These include ratings and number of reviews on Amazon, days since the book launch, and a measure of local broadband competition. The price information is missing for a few products and we include a dummy variable indicating a “missing price” in order to reduce any potential impact of these observations on the price coefficients. In the online appendix, we show our results are robust to different ways of treating missing prices.

Table 2 Main Hypotheses and Summary of Results

Hypothesis	Relevant coefficients	Prediction	Intuition	Supported?	Location
1A: Convenience	<ul style="list-style-type: none"> • Very popular products × Large bookstore entry • Very popular products × Discount store entry • Popular products × Large bookstore entry • Popular products × Discount store entry 	Negative	With an increase in the number of stores, more popular products are bought offline	Supported	Table 3, column 1, rows 1, 2, 7, 8
1B: Product selection	<ul style="list-style-type: none"> • Less popular products × Large bookstore entry • Less popular products × Discount store entry • Somewhat less popular products × Large bookstore entry • Somewhat less popular products × Discount store entry 	Negative, but less so than very popular and popular products. More negative for large bookstores than for discount stores	With an increase in the number of large bookstores, more of the less popular products are bought offline.	Not supported in the full data set	Table 3, column 1, rows 4, 5, 10, 11
2: Price	<ul style="list-style-type: none"> • Relative price × Large bookstore entry • Relative price × Discount store entry 	Positive	The impact of online discounts is tempered by local retail stores.	Supported	Table 3, column 1, rows 6, 12

the online appendix. Rows 1 and 7 of Table 3 show this most strongly: discount store and large bookstore entry decrease the likelihood of a local top 10 appearance by products in the national top 150 by 3.2 and 3.4 percentage points, respectively. These results are significant at the 1% level and economically large relative to the average likelihood that a national top 150 product appears in a local top 10 (9.8%). This suggests that online disutility costs are substantial, and changes in the distance to offline stores appear to shape consumers' channel choice.

Table 3, column 1 provides little evidence that changes in retailer distance affect the decisions of consumers to purchase less popular and somewhat less popular products. Hypothesis 1B implies that the marginal effect of store entry over the range of these products is greater for large bookstores than for discount stores. Our test of the selection effect relies on the examination of the difference between discount store and large bookstore entry. Selection implies the entry interaction coefficients for products in the 1,500–15,000 range should be more negative for large bookstores than for discount stores because they are likely to be stocked in large bookstores but not in discount stores. We do not provide evidence consistent with the selection hypothesis: the coefficients in rows 5 and 11 (or rows 4 and 10) are not significantly different from each other. We cannot separate two possible explanations for this: there is truly no effect, or, we have insufficient data to identify the effect because our "local top 10" dependent variable has relatively few unpopular products. Interestingly, in locations with universities (column 3) and in locations

with over one million people (column 4), we find support for the selection hypothesis perhaps because of more heterogeneous tastes in these locations.⁶

Figure 1 graphs the marginal effects of these interaction coefficients relative to the base of products not in the national top 15,000. It provides a visual representation of the results in Table 3, column 1, and shows that most of the impact of new store entry is found among the most popular products.

We next examine how offline store entry influences the effectiveness of online price discounts. Before discussing this interaction, we note that the negative sign in row 13 confirms the cross-price elasticity results of prior literature (e.g., Prince 2007)—price discounts increase relative sales. Hypothesis 2 conjectures that as distance to offline stores falls, online discounts become less effective. Rows 6 and 12 of Table 3 show that the coefficients on the interaction of relative price with discount stores and large bookstores are both statistically significant at the 1% level. In the absence of retailer entry, an Amazon discount relative to list price has a coefficient of -0.0237 (row 13). In contrast, when a discount store enters, this effect reduces to -0.0090 (row 6 plus row 13) and when a large bookstore enters it reduces to -0.0054 (row 12 plus row 13). Lower transportation costs are associated with less sensitivity to online discounts. Because Amazon discounts best-selling products most heavily, this means that new store entry is associated with a shift away from

⁶ We do not emphasize this result because we cannot rule out alternative explanations such as endogenous (and changing) product selection by offline retailers.

Table 3 Main Results—Difference in Difference on Store Entry

Row		(1)	(2)	(3)	(4)	(5)
		Main results	20 miles	Locations with a university	Locations over one million population	Five-month lag
<i>DiscountStoreEntry</i> (interactions)						
1	Very popular products ^a (top 150 nationally)	−0.0320 (0.0012)**	−0.0372 (0.0008)**	−0.0348 (0.0016)**	−0.0299 (0.0023)**	−0.0538 (0.0054)**
2	Popular products (151–500 nationally)	−0.0034 (0.0008)**	−0.0061 (0.0005)**	−0.0050 (0.0010)**	0.0015 (0.0016)	−0.0129 (0.0049)**
3	Moderately popular products (501–1,500 nationally)	−0.0060 (0.0006)**	−0.0080 (0.0003)**	−0.0075 (0.0008)**	−0.0023 (0.0013) ⁺	−0.0156 (0.0047)**
4	Somewhat less popular products (1,501–5,000 nationally)	−0.0082 (0.0009)**	−0.0084 (0.0004)**	−0.0087 (0.0011)**	−0.0088 (0.0019)**	−0.0195 (0.0057)**
5	Less popular products (5,001–15,000 nationally)	−0.0019 (0.0007)**	−0.0020 (0.0003)**	−0.0023 (0.0009)**	0.0014 (0.0016)	−0.0115 (0.0058)*
6	Relative Price	0.0147 (0.0022)**	0.0107 (0.0010)**	0.0158 (0.0028)**	0.0240 (0.0046)**	0.0326 (0.0089)**
<i>LargeBookstoreEntry</i> (interactions)						
7	Very popular products (top 150 nationally)	−0.0339 (0.0025)**	−0.0343 (0.0011)**	−0.0427 (0.0039)**	−0.0388 (0.0045)**	−0.0298 (0.0059)**
8	Popular products (151–500 nationally)	−0.0029 (0.0020)	−0.0045 (0.0008)**	−0.0119 (0.0027)**	−0.0073 (0.0034)*	−0.0020 (0.0044)
9	Moderately popular products (501–1,500 nationally)	−0.0022 (0.0016)	−0.0047 (0.0006)**	−0.0097 (0.0021)**	−0.0061 (0.0026)*	−0.0067 (0.0044)
10	Somewhat less popular products (1,501–5,000 nationally)	−0.0074 (0.0025)**	−0.0067 (0.0009)**	−0.0141 (0.0038)**	−0.0105 (0.0040)**	−0.0164 (0.0114)
11	Less popular products (5,001–15,000 nationally)	−0.0023 (0.0018)	−0.0022 (0.0006)**	−0.0043 (0.0019)*	−0.0049 (0.0024)*	−0.0041 (0.0056)
12	Relative price	0.0183 (0.0061)**	0.0145 (0.0021)**	0.0130 (0.0078) ⁺	0.0041 (0.0090)	−0.0111 (0.0184)
13	Other Relative price (not interacted)	−0.0237 (0.0007)**	−0.0268 (0.0008)**	−0.0259 (0.0012)**	−0.0276 (0.0015)**	−0.0226 (0.0007)**
Observations		4,051,254	4,051,254	1,808,337	1,108,643	4,051,254
Number of fixed effects		978,611	978,611	436,447	268,775	978,611
<i>R</i> -squared (with fixed effects)		0.81	0.81	0.81	0.81	0.81
<i>R</i> -squared (within)		0.07	0.07	0.07	0.08	0.07
<i>R</i> -squared (overall)		0.06	0.06	0.06	0.06	0.06
Controls						
—Dummy for missing price information		—Dummy for missing elapsed date information				
—Average rating		—Log(number of reviews)				
—Log(days since launch)		—Discount store entry				
—Time dummies		—Large bookstore entry				
—Book popularity spline		—Product-location fixed effects (differenced out)				
—Broadband competition						

Notes. Robust standard errors are in parentheses and are clustered by location time. Regressions include location-product fixed effects. *R*-squared (with fixed effects) includes fixed effects in *R*-squared computation; overall *R*-squared excludes fixed effects in *R*-squared computation; within *R*-squared centers dependent and independent variables before *R*-squared computation. For columns 1, 3, 4, and 5, we use entry in a 5.4 mile radius.

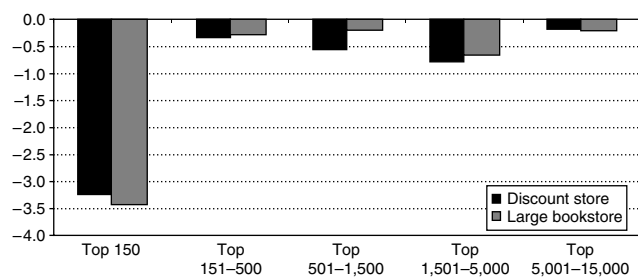
^aBase is *unpopular products* ranked 15,000 and up.

⁺Significant at the 10% level; *significant at the 5% level; **significant at the 1% level.

popular products due to both convenience and price effects.

These results are robust to a variety of different specifications (shown in the online appendix), including different distance measures, different definitions of the choice set, different definitions of the timing of entry, a different definition of broadband diffusion, different location growth rates, location-specific time

trends, different ways of treating missing prices, and different ways to define popular products including *USA Today's* bestsellers list and the *New York Times* bestsellers list. In column 5 of Table 3 we show that our results do not solely reflect short-run changes to consumer behavior after store entry; even when using a five-month lag on local store entry our qualitative results remain the same.

Figure 1 Marginal Effects from Baseline Regression (Based on Table 3, Column 1)

6. Discussion

Our results provide empirical support for the assumptions of a widely used theoretical modeling framework: spatial differentiation models that include a direct channel. We find that characteristics of these models such as offline transportation cost, online shopping disutility cost, and the prices of online and offline retailers interact to determine consumers channel choice in a way that is consistent with these models. Moreover, our results are suggestive about the relative magnitudes of some of these parameters, showing that online disutility costs can be large, even for products such as books for which nondigital attributes are relatively unimportant. Knowledge of the relative magnitudes of these parameters is important for determining the relative profitability of online and offline retailers (Balasubramanian 1998) and for determining the attractiveness of entry into the online market for incumbent offline retailers and new entrants (Liu et al. 2006, Cheng and Nault 2007).

Our empirical results also identified a set of potentially useful extensions to these models. In particular, our results suggest the usefulness of (i) understanding when the wider product availability in online stores can act as a deterrent to offline and online entry, (ii) incorporating the effect of offline transportation costs in making optimal product assortment decisions, and (iii) incorporating the effect of product popularity in modeling the impact of product returns on retailers' pricing decisions because the costs of returns to retailers and to consumers are likely to vary by product popularity and distance to stores.

Managers can also learn from our findings. For online retailers, we show how consumers' use of the online channel varies across locations. If consumers use Internet channels primarily to obtain lower prices for or more convenient access to very popular commodity products, then the expansion of large discount retailers such as Wal-Mart into new locations will result in a long-run shift in buying patterns away from the most popular products at online retailers. Furthermore, the presence of significant online disutility costs suggests that there is likely an upper bound on consumer migration to purchase commodity

products online. For offline retailers, our work shows that online retailers are relevant competitors. Competition depends on more than the number of local stores, it also depends on product overlap and disutility costs associated with the online channel. The following statement has direct practical relevance to policy makers: in 2005, the year of our data, the Federal Trade Commission blocked the Blockbuster-Hollywood Video merger partially on the basis that competition from the Internet was irrelevant and only the number of local retailers mattered.

As with any empirical work, the depth of our analysis is limited by our data. We only observe the top ten products in each location. Thus, although there is considerable heterogeneity in top products across locations and many observed purchases of less popular goods, we are limited in our ability to make inferences about purchases of very unpopular products. Similarly, we observe few locations with under 5,000 people and therefore cannot say much about channel substitution for the 11.6% of the population in smaller places. Also, we examine online behavior for just one product: books. Although our results are likely to be informative about products that share similar characteristics (such as toys) where the set of attributes is small and well-defined, our results may be less applicable for other retail categories that are sufficiently different from books (such as travel and financial services).

7. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at <http://mansci.journal.informs.org/>.

Acknowledgments

The authors thank Jeremy Fox, Korhan Gurkan, Ke-Wei Huang, Steven Klepper, Barrie Nault, Paul Pavlou, Jeffrey Prince, and numerous seminar participants for valuable suggestions. They also thank Rong Zheng and Ashley Tyrrel for outstanding assistance with the collection of data, and Gordon Eiland, Vice President of Strategy at Borders, for supplying the store opening data of Borders. Support was provided by a Berkman Faculty Development grant, the Networks, Electronic Commerce, and Telecommunications (NET) Institute, National Science Foundation (NSF) CAREER Award IIS-0643847, and Social Sciences and Humanities Research Council (SSHRC) Grant 538-02-1013. The authors assume responsibility for all errors.

References

- Ai, C., E. Norton. 2003. Interaction terms in logit and probit models. *Econom. Lett.* 80 123–129.
- Balasubramanian, S. 1998. Mail versus mall: A strategic analysis of competition between direct marketers and conventional retailers. *Marketing Sci.* 17 181–195.

- Baye, M., J. Morgan, P. Scholten. 2006. Information, search, and price dispersion. T. Hendershott, ed. *Handbook of Economics and Information Systems*. Elsevier Press, Amsterdam, 323–376.
- Bertrand, M., E. Duflo, S. Mullainathan. 2004. How much should we trust differences-in-differences estimates? *Quart. J. Econom.* **119**(1) 249–275.
- Brynjolfsson, E., M. Smith. 2000. Frictionless commerce? A comparison of Internet and conventional retailers. *Management Sci.* **46** 563–585.
- Brynjolfsson, E., Y. Hu, M. Rahman. 2008. Battle of the retail channels: How product selection and geography drive cross-channel competition. Working paper, Purdue University, West Lafayette, IN.
- Brynjolfsson, E., Y. Hu, M. Smith. 2003. Consumer surplus in the digital economy: Estimating the value of increased product variety. *Management Sci.* **49** 1580–1596.
- Cairncross, F., 1997. *The Death of Distance*. Harvard University Press, Cambridge, MA.
- Cheng, Z., B. R. Nault. 2007. Internet channel entry: Retail coverage and entry cost advantage. *Inform. Tech. Management* **8** 111–132.
- Chevalier, J., D. Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *J. Marketing Res.* **43** 345–354.
- Chun, S.-H., J. Kim. 2005. Pricing strategies in B2C electronic commerce: Analytical and empirical approaches. *Decision Support Systems* **40** 375–388.
- Ellison, G., S. Ellison. 2006. Internet retail demand: Taxes, geography, and online-offline competition. NBER Working Paper 12242, National Bureau of Economic Research, Cambridge, MA.
- Forman, C., A. Ghose, B. Wiesenfeld. 2008. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Inform. Systems Res.* **19**(3) 291–313.
- Ghose, A., M. Smith, R. Telang. 2006. Internet exchanges for used books: An empirical analysis of product cannibalization and welfare implications. *Inform. Systems Res.* **17**(1) 1–19.
- Ghose, A., R. Telang, R. Krishnan. 2005. Effect of electronic secondary markets on the supply chain. *J. Management Inform. Systems* **22** 91–120.
- Goolsbee, A. 2001. Competition in the computer industry: Online versus retail. *J. Indust. Econom.* **49** 487–499.
- Guo, L., Y. Liu. 2008. To restrain or to expand: Optimal retail store opening strategies in coping with manufacturer direct entry. Working paper, University of Illinois at Urbana-Champaign, Urbana.
- Hotelling, H. 1929. Stability in competition. *Econom. J.* **39** 41–57.
- Jeffers, P., B. R. Nault. 2007. Why online retail competition does not always benefit consumers. Working paper, Haskayne School of Business, University of Calgary, Calgary, Alberta, Canada.
- Lal, R., M. Sarvary. 1999. When and how is the Internet likely to decrease price competition. *Marketing Sci.* **18** 485–503.
- Liu, Y., S. Gupta, Z. Zhang. 2006. Note on self-restraint as an online entry-deterrence strategy. *Management Sci.* **52** 1799–1809.
- Moorthy, S., X. Zhang. 2007. Multichannel retailing: Bricks-and-mortar, clicks, or bricks-and-clicks? Working paper. University of Toronto, Toronto.
- Pan, X., B. Ratchford, V. Shankar. 2002. Price competition between pure play versus bricks-and-clicks e-retailers: Analytical model and empirical analysis. M. Baye, ed. *The Economics of the Internet and E-Commerce (Advances in Applied Microeconomics, Vol. 11)*. Elsevier, Amsterdam, 29–62.
- Prince, J. 2007. The beginning of online/retail competition and its origins: An application to personal computers. *Internat. J. Indust. Organ.* **25** 139–156.
- Rosenthal, M. 2005. Barnes & Noble book store vs. Amazon stocking. (August 29), <http://www.fonerbooks.com>.
- Salop, S. C. 1979. Monopolistic competition with outside goods. *Bell J. Econom.* **10** 141–156.
- Singh, V., K. Hansen, R. Blattberg. 2006. Market entry and consumer behavior: An investigation of a Wal-Mart supercenter. *Marketing Sci.* **25** 457–476.
- U.S. Census Bureau. 2007. E-Stats report. (May), <http://www.census.gov/estats>.
- Viswanathan, S. 2005. Competing across technology-differentiated channels: The impact of network externalities and switching costs. *Management Sci.* **51** 483–496.
- Wagner, M. 2003. Where's Wal-Mart? *Internet Retailer*. Retrieved September 1, 2007, <http://www.internetretailer.com/internet/marketing-conference/70624-wheres-wal-mart.html>.