JEONGHYE CHOI and DAVID R. BELL*

Offline retailers face trading area and shelf space constraints, so they offer products tailored to the needs of the majority. Consumers whose preferences are dissimilar to the majority—"preference minorities"—are underserved offline and should be more likely to shop online. The authors use sales data from Diapers.com, the leading U.S. online retailer for baby diapers, to show why geographic variation in preference minority status of target customers explains geographic variation in online sales. They find that, holding the absolute number of the target customers constant, online category sales are more than 50% higher in locations where customers suffer from preference isolation. Because customers in the preference minority face higher offline shopping costs, they are also less price sensitive. Niche brands, compared with popular brands, show even greater offline-to-online sales substitution. This greater sensitivity to preference isolation means that these brands in the tail of the long tail distribution draw a greater proportion of their total sales from high–preference minority regions. The authors conclude with a discussion of implications for online retailing research and practice.

Keywords: Internet, long tail, preference minority, retailing

Preference Minorities and the Internet

Local offline stores face trading area and space constraints, so they offer products that cater to the tastes of the local majority. Thus, when colocated consumers share preferences, their individual welfare is improved in that the local retail market offers products they want (Sinai and Waldfogel 2004). In contrast, consumers whose preferences are dissimilar to the majority—"preference minorities"—are likely to be underserved by local retailers or, perhaps, neglected altogether. In this study, we examine how online demand for a product category and the individual brands within it is generated from preference minorities.

*Jeonghye Choi is Assistant Professor, School of Business, Yonsei University (jeonghye@yonsei.ac.kr). David R. Bell is Xinmei Zhang and Yongge Dai Professor, the Wharton School, University of Pennsylvania (davidb@wharton.upenn.edu). The authors are grateful to Eric Bradlow, Ross Rizley, Christophe Van den Bulte, two anonymous Marketing Science Institute reviewers, and three anonymous *JMR* reviewers for comments and suggestions. They also thank seminar participants at the Haas School of Business (University of California, Berkeley), Harvard Business School, Kellogg School of Management (Northwestern University), Korea University Marketing Camp, Rotman School of Management (University of Toronto), Tel Aviv University, University of Connecticut School of Business, University of Iowa (Geography Department), University of Pennsylvania, Yonsei University, the 2008 Fall INFORMS Conference (Washington, DC), and the 2009 FORMS Conference at University of Texas at Dallas. Authorship is reverse alphabetical; both authors contributed equally to this article. Correspondence should be addressed to the first author. Roni Shachar served as associate editor for this article.

We use recent findings in economics and information systems to develop our conceptual framework and hypotheses. Larger markets deliver more product variety (Glaeser, Kolko, and Saiz 2001; Waldfogel 2003), and in many ways, the Internet serves as a "large market." As Sinai and Waldfogel (2004, p. 3) explain, "By agglomerating consumers into larger markets, the Internet allows locally isolated persons to benefit from product variety made available elsewhere." Conceptually, two forms of isolation affect online consumer demand. The first is *geographic* isolation from offline alternatives (i.e., physical distance) or transportation costs to offline alternatives. For example, Forman, Ghose, and Goldfarb (2009) show that increased distances to offline retailers lead to an increase in online demand for books. Brynjolfsson, Hu, and Rahman (2009) find that better access to offline alternatives depresses online demand for apparel. The second form of isolation, and our focus here, is *preference* isolation—a concept we elaborate on in the following paragraphs.

Profit-maximizing offline retailers allocate shelf space according to the Pareto, or "80/20," rule (Chen et al. 1999; Reibstein and Farris 1995), and "retail buyers favor products that provide the greatest returns to the shelf space and the merchandizing resources allotted them" (Farris, Olver, and De Kluyver 1989, p. 109). The implication is that offline retailers pay less attention to categories favored by preference minorities and offer smaller assortments in their stores in preference minority locations. We examine a small sample of local stores to demonstrate, through example, this assortment assumption (for a similar exercise, see Dukes, Geylani, and Srinivasan 2009, Table 1). Our empirical analysis relies on sales data from Diapers.com; therefore, we collect offline data for the diapers category. Table 1 summarizes diaper category space allocations and assortment sizes for three Fresh Grocer supermarkets and two Wal-Mart stores in the Philadelphia area. Both chains allocate more space to the diapers category and carry more diapers stockkeeping units (SKUs) when the proportion of households in the target customer group (i.e., households with babies) is higher.

More generally, in an area where the elderly are the majority of the population, young parents with newborns might not find a full assortment of baby diapers at local offline retailers. That is, young parents assume the status of preference minorities. Local stores may still allocate some shelf space to baby products, but if they do, the brands and variety offered will be limited (e.g., perhaps restricted to the popular SKUs of the leading brand, Pampers). This preference minority effect on assortment will hold, especially for product categories such as diapers, which are bulky and/or have relatively high shelf space–to-profit ratios (see Figure 1 and the related discussion). For parents who have specialized brand preferences, the overall product category effect is exacerbated: Because limited space is allocated to the product category overall, niche brands might represent only a small number of SKUs or perhaps be squeezed out altogether. Suppose that some parents have specialized prefer-

Table 1

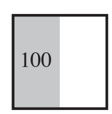
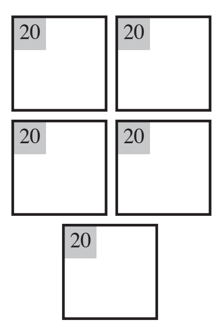
PREFERENCE MINORITIES, SHELF SPACE, AND ASSORTMENT

			Assortment: Number of SKUs				
Retailer Type	*Proportion of Households with Babies*	*Shelf Space (Width)*[1]	*Total*	*Pampers*	*Huggies*	*Luvs*	*Seventh Generation*
Fresh Grocer Supermarket							
Store 1	.106	10 ft.	28	9	15	4	0
Store 2	.155	20 ft.	50	21	19	8	2
Store 3	.163	28 ft.	63	28	25	10	0
Wal-Mart							
Store 1	.139	35 ft.	58	24	25	9	0
Store 2	.199	50 ft.	82	43	28	11	0

Notes: The retail chains we visited use the same-sized shelves across multiple locations. Within a chain, shelf height and depth are identical; thus, we provide only the width information. Each brand has potentially several variants (e.g., Pampers produces Pampers Baby Dry, Pampers Swaddlers, Pampers Swaddlers Sensitive, and Pampers Cruisers). Moreover, sizes range from "preemie" and newborns to size 6 or 7, and the number actual diapers per package can also vary.

Figure 1

PREFERENCE ISOLATION AND GEOGRAPHIC DIFFERENCES IN SHELF SPACE ALLOCATION

Target Population	*Total Population*	*Target Population as Proportion of Total*[a]	*Stores (200 sq. ft. each) and Category Shelf Space per Store*[b]	*Total Shelf Space per Market*[c]
Market A				
100	200	100/200 = 50%		100 sq. ft.
Market B				
100	1000	100/1000 = 10%		100 sq. ft.

[a]Markets A and B both have 100 residents in the target population (i.e., households with babies). Because Market B has a larger population, the target customers in Market B are, relatively speaking, preference minorities.

[b]Stores are the same size (200 sq. ft.) in both markets; however, Market B has five times as many stores because it has a larger total population (in the text, we use U.S. market data to show that whereas the number of stores increases with the population size, the size of the stores from a given chain does not). Stores allocate shelf space to categories in proportion to the size of the target market for that category (i.e., the store in Market A allocates 50%, and each store in Market B allocates 10%).

[c]The aggregate shelf space allocated to the category is the same in the two markets; however, the assortment per store will be much greater in Market A (see Table 1 and Farris, Olver, and De Kluyver 1989).

ences for chlorine-free diapers (e.g., Seventh Generation diapers), a niche product that is on average less available in local markets compared with leading brands like Pampers. These parents will have even more difficulty meeting their brand needs offline; that is, the local market characteristic of a prevalent elderly population creates preference isolation for young parents when it comes to shopping locally.

There is no absolute standard for defining minority preferences in a geographic area, so we define them by looking at the relative size of the target customer group in a local area. To implement our analysis, we construct a preference minority index (hereafter, the PM index) for each zip code equal to [1 – (Target population/Total population)] (see Forman, Ghose, and Wiesenfeld 2008; Goolsbee and Klenow 2002; Sinai and Waldfogel 2004).

Diapers.com, the largest U.S. online retailer carrying baby products, provides an excellent setting for studying geographic variation in online demand for diapers overall and for specific brands. There are several reasons the diapers category is well suited for our study. First, total diaper consumption in a location is tied to the number of babies. This means that the total number of households with babies in a location is indicative of total demand for the diapers category in that location. Second, Diapers.com carries leading national brands (Pampers, Huggies, and Luvs) and a leading niche brand (Seventh Generation) that is not available in all offline retailers. We determine exactly which offline stores in which locations carry this brand to control for region-level variation in access to popular and niche brands (for details, see the section "Data and Measures"). Third, the high shelf space–to-profit ratio for the diapers category limits product assortment in local markets for this category more than for other categories with lower shelf space–to-profit ratios (e.g., spices, vitamin pills). Fourth, the diapers category is important to offline retailers (Kumar and Leone 1988) and carried by all supermarkets.

We contribute four new substantive findings. First, we demonstrate that sales substitution, from offline retailers to online retailers, increases across local markets as the PM index increases—that is, as the relative size of the target customer group decreases. Holding the characteristics of the local environment constant, we find that online sales are higher in markets in which the target customer group is more of a preference minority. On average, online sales in "high-PM" markets (at the 90th percentile of the PM index) are more than 50% higher than in "low-PM" markets (at the 10th percentile), even though both these markets contain the same number of potential customers. Second, preference isolation reduces online price sensitivity because preference isolation implies that offline shopping costs for the category are relatively high. Our model estimates suggest that lowering online prices relative to offline prices (i.e., increasing the relative price advantage of shopping online) increases demand by approximately 30% in low-PM markets but only by approximately 10% in high-PM markets. Third, local online sales of niche brands respond more strongly to the presence of preference minorities than local online sales of "popular" brands do. High-PM markets' online sales of popular brands are approximately 40% higher than low-PM markets' sales; however, their online sales of niche brands are approximately 140% higher, even though both types of markets contain the same number of potential customers. Fourth, we find that the differential effects of preference isolation on online popular and niche brand sales have an important implication for the long tail sales distribution. Niche brands serve customers with specialized preferences and therefore typically have a lower overall sales rank, which places them in the "tail" of the long tail distribution. We find that they draw a substantially greater proportion of their total online demand from high-PM regions than popular brands do.

We organize the article as follows: The next section summarizes key ideas from the literature, introduces a conceptual framework, and describes the hypotheses. The subsequent section explains the data and measures. Then, we describe the empirical model and report and interpret the findings. We conclude with a discussion of the implications for Internet retailing theory and practice and opportunities for further research.

CONCEPTUAL FRAMEWORK AND HYPOTHESES

Demand Substitution Between Online and Offline Retailers

Online retailers, relative to offline competitors, can potentially offer consumers lower prices (Anderson et al. 2010; Brynjolfsson and Smith 2000; Goolsbee 2000), greater convenience (Balasubramanian, Konana, and Menon 2003; Forman, Ghose, and Goldfarb 2009; Keeney 1999), and more variety (Brynjolfsson, Hu, and Rahman 2009; Ghose, Smith, and Telang 2006). Among factors studied, price has received the most attention. Consumers shop online for lower prices (Brynjolfsson and Smith 2000) and to avoid local sales tax (Goolsbee 2000). Anderson et al. (2010) find that when retailers open physical stores in a location—and acquire a nexus for tax purposes—Internet sales at that location suffer because the firm must charge sales tax on Internet orders. Forman, Ghose, and Goldfarb (2009) find that when conventional booksellers enter new offline locations, Amazon.com sales at those locations decline. This increase in the convenience of the offline alternative reduces offline shopping costs and therefore reduces the attractiveness of the online alternative.

Finally, Brynjolfsson, Hu, and Rahman (2009) show that a consumer living in an area with the median number of U.S. apparel stores nearby has Internet demand that is 4.2% lower than another consumer with no offline stores nearby. They conclude that "Internet retailers face significant competition from brick-and-mortar retailers when selling mainstream products, but are virtually immune from competition when selling niche products" (p. 1755). The focal variable in their study is a measure of offline "search and transaction costs"—specifically, the number of offline stores nearby the target customers. Here, we provide additional insight by considering the preference isolation of local customers—specifically, how it reduces the Internet retailer's competition for niche products and popular products (i.e., how and why preference isolation is demand enhancing for the Internet retailer for both types of products).

Preference Isolation and Preference Minorities

Colocation of several consumers with shared needs produces two effects in a trading area. First, offline retailers pay more attention to the product category this group wants. Second, individual consumers are more able to find and buy products that suit their needs locally. This effect on local assortment is especially evident when the fixed cost of product provision is high. For example, media markets have high fixed costs, so "specialty products," such as Spanish-language programs, emerge only when sufficient numbers of customers demand them (Waldfogel 2003). Similarly, shelf space constraints make this fixed-cost argument relevant to offline retailers. Table 1 provides some preliminary evidence that shelf space allocations and assortment sizes decrease when the target customer group becomes "less important" (i.e., makes up a smaller proportion of the total market).

Figure 1 builds on that observation and conceptualizes the key relationships between preference isolation, store shelf space decisions, and the resulting assortments. It presents two hypothetical markets to show how geographic variation in preference isolation will affect online demand for a product category. Both markets have the same number of consumers in the target population (100); however, the target group makes up 50% of the consumers in Market A, but only 10% in Market B. Because Market B has a larger overall population, it contains more stores. It is well known (and perhaps obvious) that larger markets have more stores. (Christaller [1933] presents an explanation of "central place theory," a descriptive view of how the number of retail stores grows with population size.)

Stores in both markets allocate space to the product category in line with the size of the target customer population (e.g., Chen et al. 1999). Each individual store in Market B pays less attention to the target group (allocating only 10% of the store's space to the category), but both markets devote the same total amount of space (100 sq. ft.) to the product category. Critically, store-level shelf space allocation to categories according to the relative size of the target population produces more offline assortment in Market A compared with Market B, even though the total size of the target customer group and the aggregate shelf space allocated to the category are identical in both markets. The net effect is that leading brands will make it into all the stores in both markets in Figure 1 whereas niche brands will most likely be stocked only in the store in Market A (see also Farris, Olver, and De Kluyver 1989; Reibstein and Farris 1995).[1]

Before using this conceptual framework to develop our hypotheses, it makes sense to validate the three key assumptions in Figure 1 using prior research findings and market data from the 2007 U.S. Census of Business and Industry.

- *Assumption 1*: *shelf space allocations*: Prior work has assumed, either directly or indirectly, that product category space allocation in retail stores takes the relative size of the target customer group into account (e.g., Borin, Farris, and Freeland 1994; Chen et al. 1999; Murray, Talukdar, and Gosavi 2010). Table 1 also suggests this.
- *Assumption 2*: *total population and stores*: The correlation between total population and the number of local stores is strongly positive. At the Metropolitan Statistical Area (MSA) level, the correlations are .97, .86, and .96, for supermarkets, discount stores (e.g., Wal-Mart, Target), and warehouse clubs, respectively.
- *Assumption 3*: *constant store size*: Store size tends to be uniform within a retail chain. Among the 1415 Target stores in our data, for example, 99% belong to the highest size range (> 40,000 sq. ft.), and 81% belong to a single employee number range (100–249 employees).[2]

In summary, it is reasonable to assume that retailers consider the size of the target customer population when allocating space to a category and that while the number of stores increases with population, the size of individual stores in a given chain does not. Moreover, stores with less space devoted to a category focus on popular brands, so many niche brands do not "make the cut" (Anderson 1979). Thus, the amount of local product variety available offline to the target group depends on the relative size of the target customer group.

Figure 2 shows some preliminary evidence for a positive relationship between preference isolation and online sales. Recall that the PM index is equal to [1 – (target population/total population)]. Figure 2, Panel A, maps the five quintiles of the PM index in Los Angeles County, and Figure 2, Panel B, maps the cumulative number of orders per target household placed at Diapers.com for the 39 months of our data. Shading patterns show a positive correlation: In zip codes

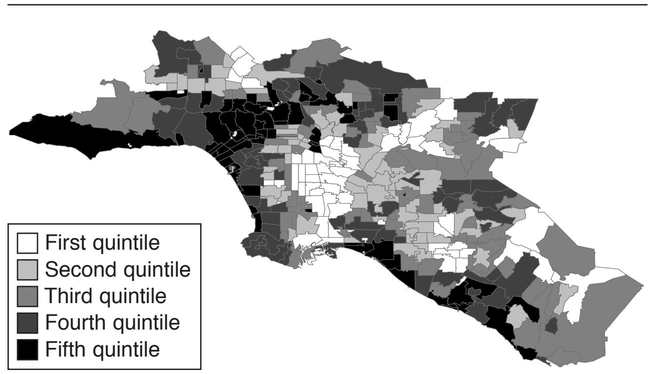
[1]Note that we implicitly assume that the five stores in Market B will offer comparable assortments focused on the popular brands—in other words, that stores will not specialize in terms of which items within a category are stocked.

[2]For retail stores, the 2007 U.S. Census of Business and Industry used 4 physical size (square feet) bins and 11 employee number bins.

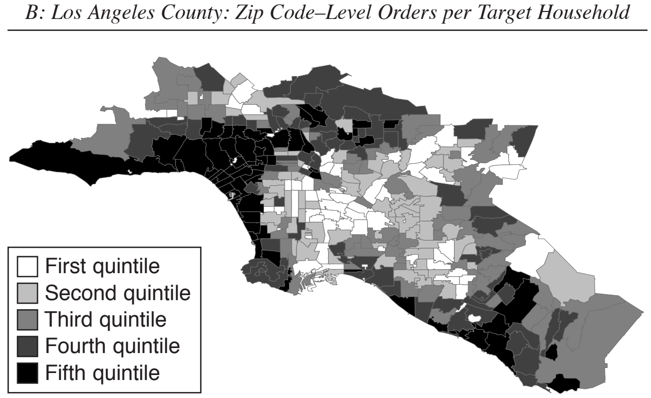
Figure 2

CORRELATION BETWEEN PRESENCE OF PREFERENCE MINORITIES AND INTERNET DEMAND

A: Los Angeles County: Zip Code–Level PM Index



B: Los Angeles County: Zip Code–Level Orders per Target Household



where households with babies are in the minority, online sales per target household are higher.

Hypotheses

Using the ideas developed in Figures 1 and 2, we specify four effects of preference isolation on geographic variation in online demand. The first two hypotheses are for overall product category effects, and the next two specify differential effects for popular and niche brands.

Category sales online (H_1). Offline retailers' category assortment decisions implicitly account for preference isolation of different customer groups. In product categories with fixed per-capita consumption in which markets with the same number of target customers have the same aggregate demand, preference isolation of the customer group is an important predictor of geographic variation in the amount of category demand satisfied online versus offline. This main prediction for the demand in the product category as a whole follows from our previous discussion and is expressed as follows:

> H_1: Category demand substitution from offline retailers to online retailers is greater in markets that have a higher PM index.

Category price sensitivity online (H_2). It would be difficult and costly to attempt to measure offline prices of all Diapers.com competitors in all geographic markets. Fortunately, prior research has suggested that at the geographic market level, sales taxes on offline purchases increase offline shopping costs and therefore increase the relative price advantage of shopping online (Anderson et al. 2010; Goolsbee 2000). Because Diapers.com prices are constant across geographies, geographic variation in the relative disparity between Diapers.com prices and local offline competitors can be captured by geographic variation in offline sales tax rates. Although the mere presence of sales taxes on offline purchases increases offline shopping costs and drives shoppers online, preference isolation suggests there will be additional geographic variation in online shoppers' price responsiveness. Preference isolation increases shopping costs in the offline market; the product category is less accessible and less well assorted. This increases the value of shopping online relative to shopping offline, holding the price disparity between the two alternatives constant. Thus, we hypothesize the following:

> H_2: Category demand in markets with a higher PM index is less sensitive to any online price advantage.

Brand sales online: popular versus niche (H_3). Offline retailers prioritize shelf space in a category; all else being equal, they stock popular brands such as the leading national brand before adding niche brands to their assortments (e.g., Farris, Olver, and De Kluyver 1989; Reibstein and Farris 1995). Preference isolation creates double jeopardy for niche brands; fewer consumers prefer them to begin with, and in high-PM markets, retailers pay even less attention to these brands. Therefore, we expect that the category-level offline-to-online substitution predicted in H_1 would be intensified for niche brands:

> H_3: Offline-to-online demand substitution for niche brands is more sensitive to geographic variation in the PM index than is offline-to-online substitution for popular brands.

Brand sales online and the long tail (H_4). Anderson (2006) popularized the long tail sales distribution concept—that is, the idea that the Internet allows sellers to stock more variety and, as a result, small percentages of sales for individual niche brands combine to contribute a large percentage of total sales and profits. Online retailers can profitably sell products that would not "make the cut" at offline retailers (Brynjolfsson, Hu, and Smith 2006). H_3 predicts that online sales of niche brands would also respond strongly to preference isolation. If H_3 holds, it follows immediately that the sales decomposition over types of geographies would differ for popular and niche brands:

> H_4: Niche brands preferred by fewer target customers and with a lower overall sales rank (i.e., those in the "tail" of the long tail) draw a greater proportion of their total online demand from high-PM regions than popular brands do.

DATA AND MEASURES

In this section, we describe our online sales data, reiterate why the diapers category is well suited to our research, and define the unit of analysis for a local market. We also describe the variables that control for geographic differences across offline markets.

Product Category

Diapers.com, the leading U.S. online retailer for baby diapers provided (1) zip code–level cumulative numbers of buyers and orders from the website's inception in January 2005 through March 2008 and (2) zip code–level cumulative sales by brand between January 2007 and March 2008. We use the three major national brands (i.e., Pampers, Huggies, and Luvs) and one niche brand that is not available in all stores (i.e., Seventh Generation) in our analysis. Seventh Generation limits distribution to bricks-and-mortar retailers that have an image of being natural or organic (e.g., Whole Foods). We control for these store locations in the empirical analysis. Furthermore, Seventh Generation is a niche brand by virtue of its appeal to a specific set of preferences; it is not simply that it is a slow seller overall.

Table 2, Panel A, presents summary statistics for the dependent variables. Orders greater than $49 (approximately 90% of all orders) qualify for free shipping and are shipped by UPS from Diapers.com warehouses. Diapers.com undertook no marketing interventions or promotional efforts targeted at preference minority regions or customers; therefore, we can assess how preference isolation in a region affects online demand there, free from explicit marketing interventions.

As we noted at the beginning of this article, diapers are especially suitable for our study. In addition to reasons given previously, because the brands are well known and shoppers face little (if any) quality uncertainty, the products have no "nondigital attributes" (Lal and Sarvary 1999). The fact that Diapers.com is a category-focused online retailer is also ideal because in our study, preference isolation is a category-level phenomenon (see Figure 1).

Unit of Analysis

The zip code is the unit of analysis; this makes sense for two reasons. First, zip codes encompass relatively self-contained groups of buyers and sellers for packaged goods

Table 2
SUMMARY STATISTICS

A: Dependent Variables: Category Buyers, Category Repeat Orders, and Brand Sales						
				Correlation		
	M	*SD*	*Mdn*	*Pampers*	*Huggies*	*Luvs*
H_1 and H_2						
Number of buyers	13.097	16.249	8			
Number of repeat orders	26.940	58.256	11			
H_3 and H_4						
Number of Pampers packages	104.409	216.504	38	—		
Number of Huggies packages	28.653	63.337	9	.815	—	
Number of Luvs packages	11.657	22.907	0	.307	.287	—
Number of Seventh Generation packages	22.936	64.446	0	.558	.561	.158

B: Independent Variables		
Variables	*M*	*SD*
Preference Isolation		
Number of total households	5620.390	3435.880
Number of households with babies aged less than six years	868.978	541.519
PM index = [1 – percentage of households with babies]	.837	.054
Online Shopping Costs		
Price: offline sales tax rate (%)	5.497	3.001
Waiting time: one-day shipping (1 = yes, 0 = no)	.190	.392
Waiting time: two-day shipping (1 = yes, 0 = no)	.199	.399
Waiting time: three-day shipping (1 = yes, 0 = no)	.327	.469
Waiting time: second warehouse led to one-day shipping (1 = yes, 0 = no)	.030	.170
Waiting time: second warehouse led to two-day shipping (1 = yes, 0 = no)	.104	.305
Convenience (H_1 and H_2): distance to nearest supermarket	1.754	2.048
Convenience (H_3 and H_4): distance to nearest supermarket selling Seventh Generation	5.145	6.730
Convenience (H_3 and H_4): distance to nearest supermarket with no Seventh Generation	1.891	2.131
Convenience: distance to nearest discount store	5.880	5.242
Convenience: distance to nearest warehouse club	11.476	12.017
Market Potential		
Local presence of stores selling baby accessories	.486	1.289
Geodemographic Controls		
Percentage of population aged 20 to 39 years	.280	.063
Percentage with bachelors and/or graduate degree	.529	.167
Percentage of female population in labor force	.556	.083
Percentage of households below the poverty line	.100	.078
Percentage of blacks	.103	.176
Percentage of apartment buildings with 50 units or more	.034	.068
Percentage of homes valued at $250,000 or more	.148	.210
Annual population growth rate from 2000 to 2004	.014	.020
Population density (thousands in square miles)	1.880	3.719

such as diapers. The most accessible offline local retail format for diapers is the local supermarket, and all zip codes that we examine have at least one supermarket. (Residential zip codes have on average four supermarkets each.) Moreover, there is roughly one discount store for every 5 zip codes and one warehouse club for every 15 zip codes. In terms of distance, supermarkets are located at approximately 2.5-mile intervals, discount stores at 8-mile intervals, and warehouse clubs at 15-mile intervals. Second, zip codes are used in many related studies of retail phenomena (for a review, see Waldfogel 2007), and following this literature, we focus on zip codes within MSAs. Metropolitan statistical areas are formed around a central urbanized area with surrounding areas that have “strong ties” to the central area. This spatial demarcation is more comprehensive than one based on geographical boundaries alone. Delaware Valley, for example, is an MSA comprising counties in Delaware, New Jersey, Maryland, and Pennsylvania. (There are 358 MSAs in the 48 contiguous states.) Limiting the analysis to zip codes within MSAs ensures that shoppers have “reasonable” travel distances to offline alternatives—in other words, they do not shop online because of a complete lack of offline stores—and is also consistent with prior research (e.g., Brynjolfsson, Hu, and Rahman 2009; Forman, Ghose, and Goldfarb 2009; Sinai and Waldfogel 2004).

Geographic Variation in Online Shopping Costs, Market Potential, and Demographics

Our hypotheses predict geographic variation in online demand as a function of geographic variation in preference isolation. Therefore, we must control for geographic variation in overall online shopping costs and other known demographic factors that affect the propensity for shoppers to buy online. Online shopping costs consist of price, waiting time, and convenience costs. Table 2, Panel B presents descriptive statistics for all independent variables.

Price. Diapers.com offers the same product prices in every zip code, but shoppers in different zip codes face different offline product prices. As we noted previously, it is not practical to gather offline prices for diapers in every

U.S. zip code; however, prior research (Anderson et al. 2010; Goolsbee 2000) has shown that geographic variation in relative offline prices for homogeneous goods can be captured using geographic variation in offline sales tax for those goods. In zip codes where offline stores collect sales tax on diapers, Diapers.com has a greater relative price advantage over offline competitors, compared with zip codes where offline taxes are not collected.[3] When offline stores collect sales tax on diapers, offline shoppers have higher shopping costs, so this should lead to greater online demand. We use data from the Department of Revenue in each state, augmented with tax-exempt information, to determine offline sales taxes for diapers. Because the diapers category is tax-exempt in some regions, we began with the relevant zip code–level tax rates (publicly available from the Department of Revenue in each state) and undertook an exhaustive manual check of local tax rates. Specifically, we made more than 1000 telephone calls to a random sample of major offline retailers and asked store employees to physically scan diaper packages and determine whether diapers were tax-exempt. The resulting data enable us to model geographic variation in disparity between the (constant) online prices and prices at offline stores.

Waiting time. Shipping times reflect online shopping convenience (Brynjolfsson and Smith 2000); thus, we need to control for geographic variation in shipping times. Shoppers are informed of the shipping time to their zip code when they order, and we collected zip code–level shipping time data from UPS. Expected shipping days under the two-warehouse regime are one to four days, and four-day shipping is the base case in the empirical model. Some zip codes saw improvements in shipping times with the launch of the second warehouse. Shorter waiting times reduce online shopping costs, and thus zip codes receiving faster shipping should produce more online demand.

Convenience. The attractiveness of shopping online is affected by the availability of offline stores (Brynjolfsson, Hu, and Rahman 2009) and the travel distance to offline stores (Forman, Ghose, and Goldfarb 2009). We used eight-digit North American Industry Classification System (NAICS) codes from the 2007 U.S. Census of Business and Industry to measure the convenience of relevant offline stores.[4] Physical distance reflects transportation costs in spatial differentiation models (e.g., Balasubramanian 1998; Bhatnagar and Ratchford 2004), so we used the actual store locations of all supermarkets, discount stores (i.e., Wal-Mart and Target), and warehouse clubs in the database and computed the expected distance to each type of store for residents of each zip code. Finally, because Pampers, Huggies, and Luvs have extensive offline distribution but Seventh Generation diapers do not, we collected location data for all stores carrying Seventh Generation and separately computed the distance from each zip code to the nearest store where this brand is available (to test H_3). Greater distances to offline retailers increase offline shopping costs by making offline shopping less convenient; this should lead to higher online demand.

Market potential and demographics. We control for several other factors likely to be correlated with zip code–level online demand for diapers. First, zip codes with greater potential for the diaper category should have more specialist retailers targeted at households with babies. To control for this, we count the number of stores selling baby accessories (e.g., Babies "R" Us) using eight-digit NAICS codes. Second, other relevant control variables are created from 2000 U.S. Census of People and Households. Measures of income and education control for the propensity to shop online and opportunity costs of time, while age, target population, and population density help control for overall market potential.

Preference Isolation and the PM Index

As we noted previously, there is no absolute determinant for minority preferences, so we focus on the proportion of households with babies aged less than six years old and define the PM index at the zip code level as [1 – (households with babies/total households)]. As we argue in the "Conceptual Framework and Hypotheses" section, geographic variation in this measure will reflect geographic variation in offline retail assortments offered to households with babies. Diapers.com, in contrast, offers the same assortment in all zip codes.

MODEL AND EMPIRICAL FINDINGS

We now describe the testing approach and the empirical findings for H_1–H_4 and also report robustness checks. One key check assesses the validity of the overall "process" argument—that is, that preference isolation is negatively correlated with offline variety in a zip code. (Recall preliminary evidence of this from Table 1; zip codes with a higher proportion of households in the target customer group [i.e., less preference isolation] have more offline variety.)

Preference Isolation and Category Sales Online

We test H_1 and H_2 with two dependent variables: (1) the number of buyers per zip code and (2) the number of repeat orders per zip code. We assume that the number of buyers (repeat orders) in zip code z in MSA m is Poisson distributed with rate parameter $\lambda_{z(m)}$. The Poisson is appropriate when the rate of occurrence is low (Agresti 2002), and in our data, the number of buyers and repeat orders in a zip code is small relative to the number of households with babies. The number of households with babies aged less than six years, $n_{z(m)}$, serves as an offset variable with its parameter constrained to one (Knorr-Held and Besag 1998; Michener and Tighe 1992).

The Poisson rate in zip code z and MSA m, $\lambda_{z(m)}$, is modeled as a function of the PM index, $PM_{z(m)}$, zip code offline sales tax rate, $TAX_{z(m)}$, the interaction of these two variables, and the other shopping cost and control variables discussed previously in the section "Data and Measures":

[3]The presence of the lowest-priced offline competitors (Wal-Mart and warehouse clubs) creates lower offline diaper prices in a zip code. Therefore, we ran additional models of category demand with a dummy variable for these stores. The dummy and its interaction with the PM index are not significant. (We provide details in the Web Appendix, Table W1, at http://www.marketingpower.com/jmraug11.) Insignificant effects in the presence of other control variables could also be due to Diapers.com striving for "Wal-Mart-level pricing" on diapers (personal communication with management).

[4]Prior research has often used six-digit NAICS codes. However, our use of eight-digit codes, while more laborious, leads to greater accuracy in store classification. As an example, using six-digit codes can lead candy stores to be grouped with supermarkets, but using eight-digit codes does not.

$$(1) \qquad y_{z(m)} \sim \text{Poisson}(\lambda_{z(m)}), \text{ and}$$

$$\begin{aligned}(2) \quad \log(\lambda_{z(m)}) &= \beta' x_{z(m)} + \varepsilon_{z(m)} \\ &= \beta_1 \times PM_{z(m)} + \beta_2 \times TAX_{z(m)} + \beta_3 \times PM_{z(m)} \\ &\times TAX_{z(m)} + \log(n_{z(m)}) + \bar{\gamma}' \times \text{Controls}_{z(m)} \\ &+ \alpha_0 + \alpha_m + \varepsilon_{z(m)},\end{aligned}$$

where $\alpha_m \sim N(0, \tau^2)$ and $\exp(\varepsilon_{z(m)}) \sim \text{Gamma}(\theta, \theta)$.

The baseline rate for regional cluster m consists of the overall baseline, α_0, and the deviation of MSA m from the overall baseline, α_m. These MSA-level random effects control for unobserved heterogeneity in the baseline rates.[5] The error term $\varepsilon_{z(m)}$ allows for overdispersion and is i.i.d. Gamma distributed with shape and scale parameters both equal to θ for identification (Cameron and Trivedi 1986; Greene 2008). After integrating over $\varepsilon_{z(m)}$, the density for $y_{z(m)}$ becomes one form of the negative binomial distribution (NBD) with mean $\mu_{z(m)}$ and variance $\mu_{z(m)}(1 + \theta^{-1}\mu_{z(m)})$. Our model has a closed-form solution up to the MSA-level random effects, and the likelihood is evaluated using numerical integration over the random effects.

Category sales online (H_1). A larger PM index indicates that the target customer group suffers from more preference isolation, which should increase the attractiveness of buying online (i.e., we expect $\beta_1 > 0$). Table 3 shows that $\beta_1 = 4.479$ ($p < .001$) for the number of buyers and $\beta_1 = 5.644$ ($p < .001$) for the number of repeat orders.[6] These estimates imply economically meaningful effects, which can be seen by comparing zip codes at different deciles on the PM index (for a similar analysis, see Sinai and Waldfogel 2004). By calculating marginal effects this way, we assume that the zip code location of the household is exogenous to any household decision to use Diapers.com (see also Forman, Goldfarb, and Greenstein 2005, p. 398).

[5]We estimate Equation 2 with MSA-level fixed effects and find qualitatively identical results. (Details are available in the Web Appendix, Table W2, at http://www.marketingpower.com/jmraug11.) The Hausman test suggests that a fixed-effects model is preferred; however, given the nearly identical estimates for the category demand model, we report random-effects results in Table 3. This is also for consistency with our brand demand models that use a multivariate NBD model with multivariate random effects (see Equation 5 and Gueorguieva 2001; Thum 1997) to parsimoniously accommodate correlations in MSA-level brand demands.

[6]We also estimate Equation 2 with the PM index as the only covariate. The β_1 estimates for the numbers of buyers and repeat orders become smaller (compare Table 3 and the Web Appendix, Table W3, at http://www.marketingpower.com/jmraug11). This suggests that the full set of control variables helps reveal the true impact of the preference isolation on online demand.

Table 3
CATEGORY-LEVEL DEMAND ESTIMATES

	Buyers		*Repeat Orders*	
	Estimate	*SE*	*Estimate*	*SE*
Intercept	–10.256*	.279	–11.563*	.536
H_1: Preference Isolation				
β_1, $PM_{z(m)}$ = [1 – percentage of households with babies]	4.479*	.306	5.644*	.593
H_2: Preference Isolation and Price Sensitivity				
β_2, $TAX_{z(m)}$ = Offline Sales Tax Rate (%)	.113*	.038	.164*	.074
β_3, $PM_{z(m)} \times TAX_{z(m)}$	–.121*	.045	–.162**	.088
Online Shopping Costs				
Waiting time: one-day shipping	.757*	.060	1.230*	.102
Waiting time: two-day shipping	.363*	.046	.568*	.078
Waiting time: three-day shipping	.213*	.041	.348*	.070
Waiting time: second warehouse led to one-day shipping	.191*	.075	.401*	.128
Waiting time: second warehouse led to two-day shipping	.128*	.053	.350*	.091
Convenience: distance to nearest supermarket	–.008*	.004	–.024*	.007
Convenience: distance to nearest discount store	.017*	.002	.021*	.003
Convenience: distance to nearest warehouse club	.005*	.001	.007*	.001
Market Potential				
Local presence of stores selling baby accessories	.007**	.004	.016*	.007
Geodemographic Controls				
Percentage of population aged 20 to 39 years	2.443*	.136	2.776*	.264
Percentage with bachelors and/or graduate degree	1.677*	.068	1.823*	.128
Percentage of female population in labor force	–.149	.128	.251	.235
Percentage of households below the poverty line	–2.551*	.179	–3.416*	.318
Percentage of blacks	–.274*	.054	–.049	.101
Percentage of apartment buildings	.722*	.104	.720*	.207
Percentage of homes valued at $250,000 or more	.856*	.047	1.640*	.094
Annual population growth rate from 2000 to 2004	10.114*	.345	10.281*	.676
Population density (thousands in square miles)	.018*	.002	.025*	.004
Variances				
θ	6.290*	.161	1.096*	.020
τ	.201*	.013	.326*	.023
–2LL	52,004		65,645	

*Significant at $p < .05$.
**Significant at $p < .10$.

For example, suppose that two zip codes have the same number of shoppers in the target population but differ in terms of total population and, therefore, on the PM index. If we compare online category demand in a low-PM market (the 10th percentile market; PM index = .79) and a high-PM market (the 90th percentile; PM index = .89), at the mean of the other covariates, this implies 6.66 buyers and 9.86 repeat orders in the low-PM market but 9.86 buyers and 16.31 repeat orders in the high-PM market. Taking trial and repeat orders together, this implies that Diapers.com sales are more than 50% higher in the high-PM market even though both markets have the same number of target consumers. Thus, our data strongly support H_1.

Category price sensitivity online (H_2). A higher offline tax rate means shoppers in the zip code have relatively higher offline shopping costs, which should increase the attractiveness of buying online (i.e., we expect $\beta_2 > 0$). Table 3 shows that $\beta_2 = .113$ ($p = .003$) for the number of buyers and $\beta_2 = .164$ ($p = .028$) for the number of repeat orders. However, H_2 is not about this straightforward main effect but rather about the interaction between preference isolation and price sensitivity. Shoppers suffering from preference isolation face higher shopping costs; thus, they need less price-based inducement to shop online. Because the Poisson/NBD model has a nonlinear functional form, we evaluated the interaction effect by computing the cross-derivative and applying the Delta method (Ai and Norton 2003) rather than simply observing the significant interaction parameters only.[7] The estimates in Table 3 imply the expected negative interaction between the PM index and the tax rate for both the number of buyers and repeat orders.

Using the estimates, we compute expected online demand by varying both the PM index and the offline sales tax rate. Assume there is a "low tax market" (the 10th percentile; no offline sales tax) and a "high tax market" (the 90th percentile; 8.25% offline sales tax rate). Higher offline sales taxes mean higher offline shopping costs, so online shopping becomes more attractive. Marginal analysis shows that when offline tax rates increase (i.e., we move from low to high tax markets), Diapers.com demand in low-PM markets increases by approximately 27%. Although an identical increase in offline shopping costs also increases Diapers.com demand in high-PM markets, it does so by only 12%. Thus, we find strong support for H_2 as well. Shoppers in high-PM markets are less sensitive to an improvement in the online price than the offline price.

Control variables. We selected control variables for online shopping costs and demographics according to prior studies, and in general, they have the expected signs. As we noted previously, higher offline tax rates reduce online shopping costs and increase demand (i.e., $\beta_2 > 0$). Less waiting time (e.g., one-day shipping vs. two-day shipping) reduces online shopping costs and increases online demand. Larger distances to offline stores increase offline shopping costs and therefore increase online demand. This intuitive effect is present and significant for discount stores (Wal-Mart and Target) and warehouse clubs. That is, we obtained our findings on the effect of preference isolation in a model that controls for "geographical isolation" (transportation costs) studied in other articles (e.g., Brynjolfsson, Hu, and Rahman 2009; Forman, Ghose, and Goldfarb 2009), along with the effects of several other control variables. Overall market potential, as indicated by the presence of baby-oriented stores, has a weaker but positive effect on online demand.[8]

Online demand also increases with known demographic drivers. Not surprisingly, Diapers.com performs better in zip codes that have higher percentages of people between 20 and 39 years of age, higher levels of educational attainment, more urban housing units, and more homes valued in excess of $250,000 or more. Less online demand in areas with higher percentages of black households and households below the poverty line is also consistent with the commonly held notion of a "digital divide" (e.g., digitaldivide.org). Finally, zip codes with greater population density and population growth have higher online demand.[9]

A less straightforward finding is the negative and significant coefficient for distance to supermarkets (though Bell and Song [2007] also observe a similar effect on online demand for a different online retailer). Greater distances to supermarkets might be expected to increase offline shopping costs and therefore increase online demand. A possible explanation for the opposite conclusion for supermarkets implied by the negative coefficient is that most shoppers will visit a supermarket to buy other categories regardless of whether they buy diapers online. Shoppers who must travel farther to offline stores and thereby incur higher offline shopping costs might try to amortize these fixed shopping costs by buying larger baskets of items per trip (Tang, Bell, and Ho 2001), which reduces their need for an online retailer. In summary, it is reassuring to observe that the hypothesized preference isolation effects hold in the presence of significant effects for other well-established control variables that account for geographic variation in online shopping costs and demographics.

Alternative Process Evidence

Our conceptual framework is built on the premise that preference isolation affects online sales by influencing the extent of product variety offered by offline retailers. We use the presence and absence of the niche brand—Seventh Generation—to provide a more direct test and shed light on the process. The logic is that, first, niche brands will be less available in high-PM markets, and second, there will be greater online demand when niche brands are less available.

[7]In a nonlinear model such as the Poisson/NBD, the sign of an interaction term is not necessarily the same as the sign of a marginal effect (Ai and Norton 2003). Moreover, Ai and Norton (2003, p. 123) report a disturbing finding: "A review of the 13 economics journals listed on JSTOR found 72 articles published between 1980 and 1999 that used interaction terms in nonlinear models. None of the studies interpreted the coefficient on the interaction term correctly."

[8]We also estimate Equation 2 without the variable for the stores selling baby accessories and obtain qualitatively identical results. Further details are available in the Web Appendix, Table W4 (http://www.marketingpower.com/jmraug11).

[9]Growing areas are more likely to have new residential buildings with young families and perhaps fewer local retailers. This negative correlation between population growth and the PM index ($r = -.254$ in our data) could inflate the coefficient for the PM index. We reestimated the models without the population growth variable and obtained qualitatively identical results. Further details are available in the Web Appendix, Table W5 (http://www.marketingpower.com/jmraug11).

To examine this, we estimate a probit model of the zip code–level presence of local stores selling Seventh Generation diapers as a function of the set of variables in Equation 2. In this equation, the estimate of the PM index is = –17.232 ($p < .001$), which implies that local stores selling Seventh Generation diapers are less likely to be present in high-PM markets. Next, we estimate the NBD model of the zip code–level online demand in Equation 2 but after replacing the PM index with a dummy variable for the presence of local stores selling Seventh Generation diapers. The dummy variable estimate is equal to –.046 ($p = .012$), which implies that online sales are lower in local markets where there are retailers carrying Seventh Generation diapers (i.e., markets with more offline variety). This two-phase empirical examination of the "process" shows evidence consistent with our underlying conceptual framework.[10]

Preference Isolation and Brand Sales Online

To test H_3, we estimated a multivariate model in which the dependent variable, $y_{i,z(m)}$, is the number of brand i diaper "standard packages" purchased in each zip code z and MSA m (i = Pampers, Huggies, Luvs, and Seventh Generation) and follows a Poisson distribution. Each diaper SKU has a different number of actual diapers, so we standardized across SKUs by converting all counts to standard units according to the most frequently purchased package sizes. The SKU "Pampers Swaddlers Jumbo Pack Size 2—80 count," for example, converts to two packages of "Pampers Swaddlers Super Mega Pack Size 2—40 count." This approach to standard units mirrors the way SKUs with multiple sizes are treated in articles using scanner panel data (e.g., Bucklin, Gupta, and Siddarth 1998). As we mentioned previously, the Poisson rate, $\lambda_{i,z(m)}$, is modeled as a function of the PM index, $PM_{z(m)}$; zip code offline sales tax rate, $TAX_{z(m)}$; the interaction of these two variables; and the other shopping cost and control variables discussed previously in the "Data and Measures" section. The number of households with babies, $n_{z(m)}$, again serves as an offset and the marginal density of $y_{z(m)}$ becomes one form of the negative binomial distribution.

$$y_{i,z(m)} \sim \text{Poisson}(\lambda_{i,z(m)}), \text{ and} \tag{3}$$

$$\begin{aligned} \log(\lambda_{i,z(m)}) &= \beta_i' x_{z(m)} + \varepsilon_{i,z(m)} \\ &= \beta_{i,1} \times PM_{z(m)} + \beta_{i,2} \times TAX_{z(m)} + \beta_{i,3} \times PM_{z(m)} \\ &\quad \times TAX_{z(m)} + \log(n_{z(m)}) + \bar{\gamma}_i' \times \text{Controls}_{z(m)} \\ &\quad + \alpha_{i,0} + \alpha_{i,m} + \varepsilon_{i,z(m)}, \end{aligned} \tag{4}$$

where $\alpha_{i,m} \sim N(0, \tau_i^2)$ and $\exp(\varepsilon_{i,z(m)}) \sim \text{Gamma}(\theta_i, \theta_i)$.

Because online demand from each of the four brands emerges from the same regional cluster, we include four random effects that follow a multivariate normal distribution (Gueorguieva 2001; Thum 1997). Joint estimation with a single model enables us to compare the effect of one covariate (e.g., the PM index) across brands.

$$\begin{pmatrix} \alpha_{\text{Pampers},m} \\ \alpha_{\text{Huggies},m} \\ \alpha_{\text{Luvs},m} \\ \alpha_{\text{SeventhGeneration},m} \end{pmatrix} \sim \text{i.i.d. MVN}\left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & r_{12}\tau_1\tau_2 & r_{13}\tau_1\tau_3 & r_{14}\tau_1\tau_4 \\ r_{12}\tau_1\tau_2 & \tau_2^2 & r_{23}\tau_2\tau_3 & r_{12}\tau_1\tau_2 \\ r_{13}\tau_1\tau_3 & r_{23}\tau_2\tau_3 & \tau_3^2 & r_{34}\tau_3\tau_4 \\ r_{14}\tau_1\tau_4 & r_{24}\tau_2\tau_4 & r_{34}\tau_3\tau_4 & \tau_4^2 \end{pmatrix}\right] \tag{5}$$

The model has a closed-form solution up to the regional random effects, and we evaluated the likelihood using numerical integration over the multivariate random effects; however, evaluation demands increase with the dimension of the random-effects vector. We alleviated this by fitting all pairwise bivariate models separately and calculating the estimates and their sampling variation for the full multivariate model (Fieuws and Verbeke 2006; Fieuws et al. 2006).

Brand sales online: popular versus niche (H_3). A larger PM index means that the target customer group suffers from more preference isolation. This not only makes it more attractive to buy the category online but also makes it especially attractive to buy niche brands online, because they suffer the most when category shelf space is reduced (Farris, Olver, and De Kluyver 1989). We expect that $\beta_1 > 0$ for all brands and that β_1 for Seventh Generation is significantly greater than the β_1 coefficients for the national brands (which are not different from one another). Table 4 shows that the estimate for Seventh Generation ($\beta_1 = 7.741$, $p < .001$) is larger than the corresponding estimates for the national brands and that these national brand estimates are not significantly different from each other. Thus, our data support H_3.

The implied quantitative effects show that online sales of the niche brand benefit disproportionately from preference isolation. At the mean of the other covariates, a high-PM market generates approximately 40% more online demand than a low-PM market for the leading brand (Pampers). The increase for the niche brand (Seventh Generation) is dramatically greater, at almost 140%, albeit from a smaller base level of sales.

Brand sales online and the long tail (H_4). Online sales of niche brands show the strongest response to preference isolation (H_3), and this immediately implies that niche brands will draw a greater proportion of their total online demand from high-PM markets. Figure 3, Panel A, is a long tail plot of expected sales for the four brands (x-axis = brands ranked by sales, and y-axis = expected sales). The three shaded bars compute expected sales from high, median, and low-PM markets. Figure 3, Panel B, shows the percentage of decomposition of the brand-specific sales across three markets. The national brands have similar decompositions and draw roughly 1.2 to 1.4 times more sales from high-PM markets than low-PM markets. The decomposition for the niche brand (Seventh Generation) shows a stark contrast. The ratio of sales from the high to low-PM markets is 48:20, or approximately 2.5:1. This elevated importance of the high-PM market for online sales of the niche brand follows

[10]We also conducted a full mediation test; the PM estimate declines slightly in size, but this change is not significant.

Table 4
BRAND-LEVEL DEMAND ESTIMATES

	Popular Brands		*Niche Brand*	
	Pampers Estimate	*Huggies Estimate*	*Luvs Estimate*	*Seventh Generation Estimate*
Intercept	–9.180*	–9.854*	–9.655*	–16.094*
H_3: Preference Isolation				
β_1, $PM_{z(m)}$ = [1 – percentage of households with babies][a]	4.521*	4.164*	3.328*	7.741*
Online Shopping Costs				
Price: $TAX_{z(m)}$ = Offline Sales Tax Rate (%)	.164*	.228*	.179	–.108
Price: $PM_{z(m)} \times TAX_{z(m)}$	–.198*	–.244*	–.201	.128
Waiting time: one-day shipping	1.035*	1.103*	.838*	1.720*
Waiting time: two-day shipping	.558*	.517*	.539*	.545*
Waiting time: three-day shipping	.280*	.387*	.483*	.404**
Waiting time: second warehouse led to one-day shipping	.403*	.394*	.522*	1.082*
Waiting time: second warehouse led to two-day shipping	.112	.416*	.195	1.081*
Convenience: distance to nearest SG store	.001	–.002	–.001	.001
Convenience: distance to nearest supermarket	–.012	–.024**	–.007	–.010
Convenience: distance to nearest discount store	.005	.020*	.004	.041*
Convenience: distance to nearest warehouse club	.008*	.008*	.008*	–.002
Market Potential				
Local presence of stores selling baby accessories	.013	.004	–.004	.013
Geodemographic Controls				
Percentage of population aged 20 to 39 years old	1.371*	.412	.392	3.915*
Percentage with bachelors and/or graduate degree	2.108*	1.950*	1.064*	3.981*
Percentage of female population in labor force	.788*	.728*	2.186*	1.369**
Percentage of households below the poverty line	–1.781*	–2.267*	–.767	–1.700
Percentage of blacks	–.721*	–.424**	–.901*	.457
Percentage of apartment buildings	2.193*	2.032*	2.245*	1.179**
Percentage of homes valued at $250,000 or more	1.667*	1.324*	–.210	.985*
Annual population growth rate from 2000 to 2004	11.503*	9.351*	6.368*	11.578*
Population density (thousands in square miles)	.014*	.022*	.012**	.017**
Variance				
θ	.724*	.427*	.210*	0.186*
τ	.344*	.309*	.300*	0.671*
r_{12} (Pampers, Huggies)	.815*			
r_{13} (Pampers, Luvs)	.307			
r_{14} (Pampers, Seventh Generation)	.558*			
r_{23} (Huggies, Luvs)	.288			
r_{24} (Huggies, Seventh Generation)	.561*			
r_{34} (Luvs, Seventh Generation)	.561*			

*Significant at $p < .05$
**Significant at $p < .10$.
[a]The estimate of the PM index for the niche brand, Seventh Generation, is significantly larger ($p < .01$) than those for the national brands, Pampers, Huggies, and Luvs, while these three estimates for the national brands are not significantly different from each other.

directly from the finding that the β_1 estimate for Seventh Generation in the multivariate model is significantly larger than the corresponding β_1 estimates for the national brands ($p < .01$). In line with H_4, preference isolation is especially conducive to online sales of niche brands.

CONCLUSION

Drawing on theory and empirical findings in economics and information systems, we introduce the concept of preference isolation as a driver of offline-to-online sales substitution in local markets. First and foremost, we conjecture and find that category-level sales substitution to online retailers is greater in markets that have a higher PM index (H_1). Furthermore, high-PM markets are less price sensitive than low-PM markets (H_2). Finally, offline-to-online demand substitution due to preference isolation is significantly greater for niche brands than for popular brands (H_3), and niche brands therefore draw a greater proportion of their total sales from high PM markets (H_4).

Implications for Online Retailing

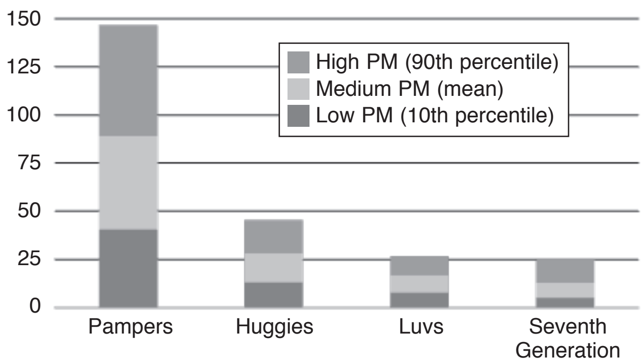
Our findings imply a new and important geographic targeting heuristic for Internet retailers. It is natural that an Internet retailer would focus on markets in which the absolute number of potential customers is high; however, customers in these markets should be relatively well served by offline retailers. Internet retailers must also consider the relative size of their target customer group in a given location. Our estimates show that online category sales in high-PM markets are at least 50% greater than they are in low-PM markets, even though both types of markets have the same total number of customers who need the category. Selling niche brands in high-PM markets is especially attractive because these customers face high offline shopping costs and are therefore less price sensitive.

Offline retailers can improve the economics of stocking slower-moving SKUs (and therefore increase the variety they offer) by using distributors who stock in less-than-case pack-out quantities. Even so, there are several kinds of

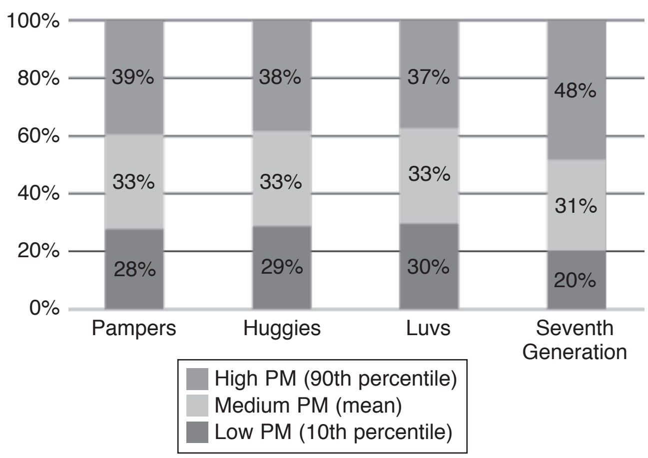
Figure 3

THE CONTRIBUTION OF LOCAL MARKETS TO BRAND SALES ONLINE

A: The Long Tail Sales Distribution[a]



B: Decomposition of Brand Sales by Market Type[b]



[a]We fix the PM index at the 10th percentile, the mean, and the 90th percentile. Holding everything else equal, we compute the expected sales for each brand in each of the three prototypical markets.

[b]Preference isolation leads all brands to have higher online sales in markets with higher PM indexes (H_1). For the niche brand (Seventh Generation), substitution from offline to online sales is even stronger (H_3). Thus, preference isolation implies the greatest *relative* proportion of sales that come from high-PM markets, and the effect is amplified for niche brand (H_4). The niche brand sales decomposition in high- versus low-PM markets is 48:20, or approximately 2.5:1.

bulky or low-value categories or niche brands that need minimum facing and are difficult for offline retailers to justify. Online retailers can exploit this assortment gap at offline stores, particularly in high-PM markets. We demonstrate this phenomenon using the diapers category, but we believe that other categories with similar properties should also benefit in the same way.

Our study adds to evidence that consumer benefits from online shopping are contextual and need to be assessed relative to offline options (Anderson et al. 2010; Brynjolfsson, Hu, and Rahman 2009; Choi, Hui, and Bell 2010; Forman, Ghose, and Goldfarb 2009). In particular, we show how a specific form of consumer isolation—preference isolation—explains geographic variation in online demand. The net benefit to individual consumers from online shopping depends on not only where they live but also who lives next to them.

Limitations and Further Research

We bring the concept of preference isolation to the substantive marketing problem of how online demand derives from an assortment deficiency in the offline market. Many more opportunities exist to develop theory and empirical analysis for how the availability of an Internet option affects behavior. Agrawal and Golfarb (2008) show that the BITNET reduced “academic isolation” and thereby facilitated an approximately 40% increase in multi-institutional research collaboration among engineering faculty. Consumer-to-consumer interaction is having similar dramatic effects on shopping behavior, and new business forms such as Gilt.com, Groupon.com, and Yipit.com are emerging as a result. These new institutions are certainly worthy of formal analysis.

Our research implies the possibility of endogenous preference for variety: Customers in the preference minority might go online for the reasons we suggest (H_1) but, having got there, expand their brand preferences within a category. In our data, preferences seem stable, in that there is little switching among brands, conditional on shopping at Diapers.com, but this need not be true in other product categories (e.g., Overby and Jap 2009). Finally, greater neighborhood diversity implies the presence of more preference minorities in more product categories. We are currently working with data from an Internet retailer specializing in a men’s clothing and accessories brand and have found a positive cross-sectional correlation between region-level demand and the so-called ESRI diversity index. This index represents the “likelihood that two persons, chosen at random from the same area, belong to different races or ethnic groups” (ESRI 2010).

REFERENCES

Agrawal, Ajay and Avi Goldfarb (2008), “Restructuring Research: Communication Costs and the Democratization of University Innovation,” *American Economic Review*, 98 (4), 1578–90.

Agresti, Alan (2002), *Categorical Data Analysis*. New York: John Wiley & Sons.

Ai, Chunrong and Edward C. Norton (2003), “Interaction Terms in Logit and Probit Models,” *Economic Letters*, 80 (1), 123–29.

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

Anderson, Eric T., Nathan M. Fong, Duncan I. Simester, and Catherine E. Tucker (2010), “How Sales Taxes Affect Customer and Firm Behavior: The Role of Search on the Internet,” *Journal of Marketing Research*, 47 (April), 229–39.

Anderson, Evan E. (1979), “An Analysis of Retail Display Space: Theory and Methods,” *Journal of Business*, 52 (1), 103–118.

Balasubramanian, Sridhar (1998), “Mail Versus Mall: A Strategic Analysis of Competition Between Direct Marketers and Conventional Retailers,” *Marketing Science*, 17 (3), 181–95.

———, Prabhudev Konana, and Nirup M. Menon (2003), “Customer Satisfaction in Virtual Environments: A Study of Online Investing,” *Management Science*, 49 (7), 871–89.

Bell, David R. and Sangyoung Song (2007), “Neighborhood Effects and Trial on the Internet: Evidence from Online Grocery Retailing,” *Quantitative Marketing and Economics*, 5 (4), 361–400.

Bhatnagar, Amit and Brian T. Ratchford (2004), “A Model of Retail Format Competition for Non-Durable Goods,” *International Journal of Research in Marketing*, 21 (1), 39–59.

Borin, Norm, Paul W. Farris, and James R. Freeland (1994), "A Model for Determining Retail Product Category Assortment and Shelf Space Allocation," *Decision Sciences*, 25 (3), 359–84.

Brynjolfsson, Erik, Yu (Jeffrey) Hu, and Michael D. Smith (2006), "From Niches to Riches: The Anatomy of the Long Tail," *MIT Sloan Management Review*, 47 (4), 67–71.

——— and Michael D. Smith (2000), "Frictionless Commerce? A Comparison of Internet and Conventional Retailers," *Management Science*, 46 (4), 563–85.

———, ———, and Mohammad S. Rahman (2009), "Battle of the Retail Channels: How Product Selection and Geography Drive Cross-Channel Competition," *Management Science*, 55 (11), 1755–65.

Bucklin, Randolph E., Sunil Gupta, and S. Siddarth (1998), "Determining Segmentation in Sales Response Across Consumer Purchase Behaviors," *Journal of Marketing Research*, 35 (May), 189–97.

Cameron, A. Colin and Pravin K. Trivedi (1986), "Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests," *Journal of Applied Econometrics*, 1 (1), 29–53.

Chen, Yuxin, James D. Hess, Ronald T. Wilcox, and Z. John Zhang (1999), "Accounting Profits Versus Marketing Profits: A Relevant Metric for Category Management," *Marketing Science*, 18 (3), 208–229.

Choi, Jeonghye, Sam K. Hui, and David R. Bell (2010), "Spatio-Temporal Analysis of Imitation Behavior Across New Buyers at an Online Grocery Retailer," *Journal of Marketing Research*, 47 (February), 75–89.

Christaller, Walter (1933), *Die zentralen Orte in Suddeutschland*. Jena, Germany: Gustav Fischer. (Translated [in part] by Charlisle W. Baskin (1966), as *Central Places in Southern Germany*. Englewood Cliffs, NJ: Prentice Hall).

Dukes, Anthony J., Tansev Geylani, and Kannan Srinivasan (2009), "Strategic Assortment Reduction by a Dominant Retailer," *Marketing Science*, 28 (2), 309–319.

ESRI, "Methodology Statement: 2010 Diversity Index," white paper, (accessed May 9, 2011), [available at http://www.esri.com/library/whitepapers/pdfs/diversity-index-methodology.pdf].

Farris, Paul, James Olver, and Cornelis de Kluyver (1989), "The Relationship Between Distribution and Market Share," *Marketing Science*, 8 (2), 107–128.

Fieuws, Steffen and Geert Verbeke (2006), "Pairwise Fitting of Mixed Models for the Joint Modeling of Multivariate Longitudinal Profiles," *Biometrics*, 62 (2), 424–31.

———, ———, Filip Boen, and Christophe Delecluse (2006), "High Dimensional Multivariate Mixed Models for Binary Questionnaire Data," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 55 (4), 449–60.

Forman, Chris, Anindya Ghose, and Avi Goldfarb (2009), "Competition Between Local and Electronic Markets: How the Benefit of Buying Online Depends on Where You Live," *Management Science*, 55 (1), 47–57.

———, ———, and Batia Wiesenfeld (2008), "Examining the Relationship Between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets," *Information Systems Research*, 19 (3), 291–313.

———, Avi Goldfarb, and Shane Greenstein (2005), "How Did Location Affect Adoption of the Commercial Internet? Global Village vs. Urban Leadership," *Journal of Urban Economics*, 58 (3), 389–420.

Ghose, Anindya, Michael D. Smith, and Rahul Telang (2006), "Internet Exchanges for Used Books: An Empirical Analysis of Product Cannibalization and Welfare Impact," *Information Systems Research*, 17 (1), 3–19.

Glaeser, Edward L., Jed Kolko, and Albert Saiz (2001), "Consumer City," *Journal of Economic Geography*, 1 (1), 27–50.

Goolsbee, Austan (2000), "In a World Without Borders: The Impact of Taxes on Internet Commerce," *Quarterly Journal of Economics*, 115 (2), 561–76

——— and Peter J. Klenow (2002), "Evidence of Learning and Network Externalities in the Diffusion of Home Computers," *Journal of Law and Economics*, 45 (2), 317–43.

Greene, William (2008), *Econometric Analysis*. Upper Saddle River, NJ: Prentice Hall.

Gueorguieva, Ralitza (2001), "A Multivariate Generalized Linear Mixed Model for Joint Modeling of Clustered Outcomes in the Exponential Family," *Statistical Modeling*, 1, 177–93.

Keeney, Ralph L. (1999), "The Value of Internet Commerce to the Customer," *Management Science*, 45 (4), 533–42.

Knorr-Held, Leonhard and Julian Besag (1998), "Modeling Risk from Disease in Time and Space," *Statistics in Medicine*, 17 (18), 2045–2060.

Kumar, V. and Robert P. Leone (1988), "Measuring the Effect of Retail Store Promotions on Brand and Store Substitution," *Journal of Marketing Research*, 25 (May), 178–85.

Lal, Rajiv and Miklos Sarvary (1999), "When and How Is the Internet Likely to Decrease Price Competition?" *Marketing Science*, 18 (4), 485–503.

Michener, Ron and Carla Tighe (1992), "A Poisson Regression Model of Highway Fatalities," *American Economic Review*, 82 (2), 452–56.

Murray, Chase C., Debabrata Talukdar, and Abhijit Gosavi (2010), "Joint Optimization of Product Price, Display Orientation and Shelf-Space Allocation in Retail Category Management," *Journal of Retailing*, 86 (2), 125–36.

Overby, Eric and Sandy Jap (2009), "Electronic and Physical Market Channels: A Multi-Year Investigation in a Market for Products of Uncertain Quality," *Management Science*, 55 (6), 940–57.

Reibstein, David J. and Paul W. Farris (1995), "Market Share and Distribution: A Generalization, a Speculation, and Some Implications," *Marketing Science*, 14 (3), 190–202.

Sinai, Todd and Joel Waldfogel (2004), "Geography and the Internet: Is the Internet a Substitute or a Complement for Cities?" *Journal of Urban Economics*, 56 (1), 1–24.

Tang, Christopher S., David R. Bell, and Teck-Hua Ho (2001), "Store Choice and Shopping Behavior: How Price Format Works," *California Management Review,* 43 (2), 56–74.

Thum, Yeow Meng (1997), "Hierarchical Linear Models for Multivariate Outcomes," *Journal of Educational and Behavioral Statistics*, 22 (1), 77–108.

Waldfogel, Joel (2003), "Preference Externalities: An Empirical Study of Who Benefits Whom in Differentiated-Product Markets," *RAND Journal of Economics*, 34 (3), 557–68.

——— (2007), *The Tyranny of the Market: Why You Can't Always Get What You Want*. Cambridge, MA: Harvard University Press.