PRADEEP K. CHINTAGUNTA and JEAN-PIERRE DUBÉ*

The marketing literature has addressed the issues of heterogeneity and endogeneity when estimating a choice model with household-level panel data. When using these data at the stockkeeping unit or the Universal Product Code level, choices for each item in each of the time periods under consideration cannot be observed. Without such information, it is difficult to control for item- and time period–specific unmeasured characteristics because there is no information on alternatives during those periods in which they are not purchased by any of the panelists. In general, when a product category has many alternatives, each with fairly small shares, the household sample may not contain sufficient choices for each alternative, thus negatively affecting the ability to control for endogeneity with household data. In contrast, because aggregate store-level data (for those stores in which the panel makes purchases) are the aggregation of purchases by all households visiting the stores, the data contain the time period–specific item-level information required to account for endogeneity as long as each item has some sales in each time period. Given the relative merits of household data to estimate the distribution of heterogeneity and store-level data to address the endogeneity problem, the authors propose an integrated estimation procedure that uses the information in both sources. They provide empirical results from their model using data on the fabric softener market. They extend their approach to situations in which there is variation in purchase quantities that households choose.

Estimating a Stockkeeping-Unit-Level Brand Choice Model That Combines Household Panel Data and Store Data

Accounting for heterogeneity across households or consumers when studying the effects of marketing activities on choice behavior is an important area of marketing research. Allowing for heterogeneity provides a more realistic portrayal of consumer choice behavior, and failure to control for it can result in biases in the estimated mean responses to marketing variables (e.g., Allenby and Rossi 1999; Chintagunta, Jain, and Vilcassim 1991; Kamakura and Russell 1989). Although a majority of research on understanding the nature of the heterogeneity distribution has involved the use of household scanner data, recent research in marketing has also attempted to estimate the distribution of heterogeneity from store-level data (e.g., Besanko, Dubé, and Gupta 2003). Because choices (in store data) are aggregated across the population of shoppers, within-panel information is lost, and thus it is not possible to control for the dependence across a given household's choices over time. Instead, heterogeneity is inferred from asymmetries in relative market share co-movements across weeks in response to observed price variation. Effectively, heterogeneity in store data is identified from non-independence-from-irrelevant-alternatives (IIA) aggregate substitution patterns. Thus, although accounting for heterogeneity may be important with store data, it is not possible to test formally an aggregate model with heterogeneity against an alternative model with non-IIA substitution patterns. Therefore, from the perspective of understanding heterogeneity, household panel data are superior to store data for learning about household-specific information.

*Pradeep K. Chintagunta is Robert Law Professor of Marketing (e-mail: pradeep.chintagunta@gsb.uchicago.edu), and Jean-Pierre Dubé is Associate Professor of Marketing (e-mail: jdube@gsb.uchicago.edu), Graduate School of Business, University of Chicago. The authors thank Greg Allenby, Pete Fader, and K. Sudhir for helpful comments and discussion. The authors are grateful to the Kilts Center for Marketing at the University of Chicago for financial support. Chintagunta is indebted to the Harvard Business School, where most of this article was written, and Dubé acknowledges the True North faculty fund for research support.

More recently, in addition to heterogeneity, the literature has identified another source of bias in the estimation of choice models. The bias stems from item or brand characteristics, which influence consumer choices but are unobserved by the researcher. Because they are unobserved, these characteristics form an additional error component in the choice model that varies by item and across time periods but not across households. This term captures dependence across households' choices in a given week. Because this is an aggregate shock, estimation problems could arise if the marketer observes this unobserved brand characteristic and incorporates it into marketing decision making (e.g., price setting). If this error term is correlated with the marketing variables included in the model, ignoring it can bias the mean responses to marketing variables; this is the so-called endogeneity problem.

One solution to this problem is to add more structure and to write out the joint likelihood of the choices and the prices (e.g., Villas-Boas and Zhao 2005; Yang, Chen, and Allenby 2003). A drawback of this approach is that if no choices of an item are observed in a particular time period, it is difficult to infer the unobserved item characteristic for that item in that week. This situation is particularly acute if there is interest in studying choice behavior at the stockkeeping unit (SKU)/Universal Product Code (UPC)/item level (see Fader and Hardie 1996). A plot of the weekly unit sales based on aggregating the purchases from the household panel for the best-selling UPC in the fabric softener data set (one of the data sets we use in our empirical application) clearly indicates that in several weeks, the panel does not collectively make any purchases of this UPC. Because in this case the unobserved attributes are potentially UPC and week specific, we have no information from the household panel to learn about these unobserved factors if we have access only to household data.

As Berry (1994) outlines, an alternative solution to this problem requires computing the mean utility across households of each brand each week when fitting the choice model. The mean utility is the linear sum of the mean intrinsic preference for that brand, the mean effects of marketing activities, and the unobserved (to the econometrician) brand characteristics, where the correlation between the marketing activities (specifically, price) and the unobserved characteristics generates the endogeneity problem. Berry suggests extracting the information about the mean utility directly from the products' aggregate shares and then recovering the mean preferences and marketing effects by regressing these mean utilities on the marketing activities through an instrumental variables (IV) regression. Because the aggregate share is a function of the mean utilities, the latter can be obtained by "inverting" the share function. Thus, the aggregate weekly shares of each brand in the stores in which the household panel makes purchases (precisely the information available in store data) are required.

Instead of using store data, we could estimate the weekly mean utilities directly as weekly brand intercepts using household panel data (Chintagunta, Dubé, and Goh 2005). Alternatively, we could apply Berry's (1994) approach to household data by approximating the weekly shares using the aggregated choices of each brand in each week across all households in the panel (Goolsbee and Petrin 2003). However, typical scanner panels may be too sparse to provide a reasonable sample estimate of the shares of each product each week or an intercept for each brand each week. Because we are interested in estimating a choice model at the SKU or item level, the aggregated shares from the panel are likely to be zero for many items in several weeks.

Given the relative merits of using household data to recover the heterogeneity distribution and store-level data to address the endogeneity problem in situations in which there is little or no information available in the household data for this purpose, a natural question is whether both household and store data can be combined to obtain a single set of price elasticity estimates. Rather than obtaining two different sets of estimates from the two data sources, we propose a method that provides one set of estimates using information from the household and store data sets. In particular, we are interested in (1) estimating the mean effects of marketing activities, (2) accounting for endogeneity using information from the store data, and (3) obtaining the distribution of heterogeneity across households using information from the household data. Essentially, the idea is to offset the drawbacks of one source with information from the other and combine the best aspects of both data. The key assumption that we make in our analysis is that store data provide information on the item-level unmeasured characteristics in each time period and that household data provide an accurate representation of the heterogeneity distribution. We combine the two sources into an integrated estimation procedure because store-level data are an aggregation of the choices of the population of households shopping in the store in a given week (Gupta et al. 1996). At the same time, the household data are a random sample of households from this population. The main reason for combining the household and aggregate databases is to obtain unbiased estimates of the structural parameters of the model (i.e., consumer tastes). As Bronnenberg, Rossi, and Vilcassim (2005) discuss, the consistent estimation of structural parameters is crucial when the estimates are used for policy simulations.

In essence, the motivation for our approach to combining store and household data is as follows: When household data are sparse (e.g., for infrequently purchased product categories or when the data are at the SKU level), it is difficult to address the endogeneity issue because there is no information available on the unobserved characteristics for the items during certain weeks. Similarly, store data are limited in their ability to recover the distribution of heterogeneity across consumers (for conditions under which this is feasible, see Bodapati and Gupta 2004); household data are required for this purpose, especially if the objective is to exploit the distribution for targeting purposes. Furthermore, it is not possible to account for the purchase quantity decision at the household level using only store data.

We organize the remainder of this article as follows: In the next two sections, we discuss our model and estimation procedure, respectively. We then discuss the SKU-level data from the fabric softeners category that are sparse at the household level. In the subsequent section, we provide the empirical results and note their substantive implications. We also extend our approach to incorporate purchase quantities (besides purchase incidence and brand choice) and to provide the empirical results from the orange juice category.

MODEL

Suppose there are J items (or SKUs) in the category and the J + 1th no-purchase option. We describe a product j by a time-varying M × 1 vector of marketing variables (price, feature advertisement, and display), X_{jt}; a time-invariant K × 1 vector of product characteristics, A_j; and an additional attribute, ζ_{jt}, to control for product characteristics that are observed by the consumer but not by the researcher. On a trip during week t, household h has the familiar conditional indirect utility for alternative j:

$$(1) \qquad U_{hjt} = A_j\alpha_h + X_{jt}\beta_h + \zeta_{jt} + \varepsilon_{hjt},$$

where β_h is the household's response to marketing variables, α_h is the household's tastes for product characteristics, and ε_{hjt} is an extreme value error term. Note that we normalize the systematic component of indirect utility of the outside option to be zero, $U_{h(J+1)t} = \varepsilon_{h(J+1)t}$.

We assume the vector of taste parameters is distributed multivariate normal: $\{\beta_h, \alpha_h\} \sim N(\{\beta, \alpha\}, \Sigma)$, where $\{\beta, \alpha\}$ is the mean vector, and Σ is the covariance matrix. As Berry, Levinsohn, and Pakes (hereinafter BLP) (1995) discuss, households that purchase item j will switch to a product with similar characteristics when the price of item j rises. This result is due to the households having attribute tastes (α_{hk} for attribute k) that differ from the mean levels. This provides flexible substitution patterns due to the heterogeneity in tastes for attributes. In addition, we explicitly allow for correlations in preferences and between preferences and price sensitivities. This provides further flexibility to the model. We can rewrite the indirect utility for product j as

$$(2)\ U_{hjt} = A_j\alpha + X_{jt}\beta + \zeta_{jt} + \sum_{k=1}^{K} A_{jk}v_{\alpha hk} + \sum_{m=1}^{M} X_{jtm}v_{\beta hm} + \varepsilon_{hjt},$$

where the terms v are normal random variables drawn from a multivariate normal distribution with mean vector 0 and covariance matrix Σ (as we described previously).

In this model, the role of the unobserved attribute, ζ_{jt}, is particularly critical because we do not have an item-specific intercept to control for all time-invariant aspects of quality. A potential concern for the researcher is that if retailers set prices strategically and they observe ζ_{jt}, prices would likely be correlated with ζ_{jt} both over time and across products. If this correlation were positive, the researcher would underestimate the consumer sensitivity to prices. Such price endogeneity is not alleviated simply by resorting to individual data (Chintagunta, Dubé, and Goh 2005; Goolsbee and Petrin 2003). To resolve this problem, we use weekly item-specific intercepts, $\lambda_{jt} = A_j\alpha + X_{jt}\beta + \zeta_{jt}$, to capture the component of utility that is common across all households at time t, including the unobserved attribute. This term is frequently referred to as the "mean utility" because it is the average utility across consumers at time t at the true parameter values. In the estimation section, we discuss how we estimate these intercepts. To simplify the notation, we also define the following:

$$\mu_{hjt}(A_j, X_{jt}, v, \Sigma) = \sum_{k=1}^{K} A_{jk}v_{\alpha hk} + \sum_{m=1}^{M} X_{jtm}v_{\beta hm}.$$

We obtain the probability that consumer h chooses brand j during week t by integrating out the extreme value error term to obtain

$$(3)\ P_{hjt}(A, X_t, \lambda_t, v, \Sigma) = \frac{\exp[\lambda_{jt} + \mu_{hjt}(A_j, X_{jt}, v, \Sigma)]}{1 + \sum_{i=1}^{J} \exp[\lambda_{it} + \mu_{hit}(A_i, X_{it}, v, \Sigma)]}, \quad j = 1, ..., J,$$

where $A = (A_1, A_2, ..., A_J)'$, $X_t = (X_{1t}, ..., X_{Jt})'$, and $\lambda_t = (\lambda_{1t}, ..., \lambda_{Jt})'$. After integrating out the unobserved heterogeneity, the density of a household's observed sequence of choices over time is given by the following:

$$(4)\ L_h(Y_h|A, X, \lambda; \Sigma) = \int \prod_{t=1}^{T_h} \prod_{j=1}^{J} P_{hjt}(A, X_t, \lambda_t, \Sigma, v)^{Y_{hjt}} \phi(v)\partial v,$$

where $Y_h = (Y_{1h}, ..., Y_{Th})'$, $Y_{hjt} = 1$ if brand j is chosen on trip t, and $\phi(\cdot)$ is the density of the multivariate normal distribution. Note that the estimation of λ and Σ is not biased by price endogeneity because we effectively condition on ζ, which is accounted for by λ.

We can address the correlation between prices and ζ by modeling a specific form of pricing "game" that, in equilibrium, generates a likelihood for the observed shelf prices (Villas-Boas and Zhao 2005; Yang, Chen, and Allenby 2003). The additional imposed structure could improve the precision of the estimated demand parameters if the game is specified correctly. However, misspecification of the pricing game leads to biases in the estimated demand parameters. Furthermore, pricing games with large differentiated product demand systems, such as those that we study, suffer from multiple equilibriums (i.e., multiple price vectors that could satisfy the optimal price conditions for a given demand system). In such cases, the likelihood is not well defined (see the published discussion accompanying Yang, Chen, and Allenby 2003). For these reasons, we can consider our approach more robust.

At the store level, we define the market size as H_t, whereby a continuum of consumers of mass H_t visits the store in week t. The error terms ε in Equation 1 are independently distributed over this continuum.[1] We define the set of consumers in week t who choose brand j (i.e., brand j's "market segment") as $B_{jt} = \{(v_h, \varepsilon_h): U_{hjt} \geq U_{hjk}, k = 1, ..., J+1\}$. Then, the market share of brand j is given by the following:

$$(5) \qquad S_{jt} = \int_{B_{jt}} \phi(v)g(\varepsilon)\partial v \partial \varepsilon, \ j = 1, ..., J+1,$$

where $g(\varepsilon)$ is the extreme value density. Because we can integrate over the extreme value density analytically, we can rewrite the market share as

$$(6) \qquad S_{jt} = \int P_{jt}(A, X_t, \lambda_t, \Sigma, v)\phi(v)\partial v, \ j = 1, ..., J+1.$$

[1]We can consider the observed cross-section of households in our panel data a finite number of random draws from this continuum.

Conditional on ζ, the market share in Equation 6 is deterministic, and the aggregate demand for brand j is $Q_{jt} = H_tS_{jt}$.[2] At the true values of λ_{jt} and Σ, this equation will hold exactly, conditional on the data $(A_t, X_t)'$. Thus, the model would exactly predict the observed aggregate store shares. This outcome is due to our effective conditioning on the aggregate error term, ζ, which is contained in the term λ. In the subsequent section, we combine the systems in Equations 4 and 6 to leverage both our household- and our store-level data.

ESTIMATION PROCEDURE

Our objective is to estimate the corresponding taste parameter vectors, α, β, and Σ, which are the means and variances, respectively, of our normally distributed random coefficients specification. We now explain how we use the household model in conjunction with the corresponding store-level model to recover our model parameters while controlling for price endogeneity. As we discussed in the previous section, we resolve the endogeneity problem by including the weekly item-specific intercepts, λ. However, the estimation of these intercepts poses an econometric challenge. Note that we are not inherently interested in the values of λ per se, but we use them to recover the mean taste parameters, α and β.

In theory, we could estimate the intercepts, λ, directly using maximum likelihood (Chintagunta, Dubé, and Goh 2005). In practice, however, we would have too many parameters to manage computationally, especially when considering choices at the SKU level. Instead, we use a result from the work of Berry (1994), who shows that for given values of Σ, there exists a unique set of intercepts, λ_t, such that the predicted shares $S_{jt}(A, X_t, \lambda_t, \Sigma)$ are matched exactly to the observed shares, S_{jt}. Thus, we can search numerically for values of λ that match Equation 6 exactly to the observed aggregate shares.[3] These values are functions $\lambda_{jt}(S_{jt}, \Sigma)$. Unlike Goolsbee and Petrin's (2003) method, we match Equation 6 to the weekly aggregate store shares, S_{jt}, for those stores in which the households shop rather than to the weekly household sample shares. The use of the actual store aggregate shares versus the household sample aggregate shares has several advantages. The household sample shares are stochastic, distributed multinomial Dirichlet due to sampling error. This sampling error generates additional uncertainty around the parameter estimates. In addition, at the SKU level, there are frequently weeks during which some of the alternatives are not chosen by any of the sample households, and thus a sample share cannot be computed. By using the store-level shares, we do not need to worry about the impact of sampling error at the household panel level on the estimation of Σ or the computation of λ, and we do not need to worry about feasibility constraints due to sparseness in the number of observed purchases of a brand in any week.[4] For these reasons, the store-level data are more appropriate for learning about the mean utility, λ.

To estimate the standard deviations in the covariance matrix, Σ, we substitute $\lambda_{jt}(S_{jt}, \Sigma)$ back into the density of consumers' choice histories:

$$(7) \qquad L_h(Y_h|A, X; \Sigma)$$

$$= \int_{-\infty}^{\infty}\prod_{t=1}^{T_h}\prod_{j=1}^{J} P_{hjt}\left[A, X_t, \lambda_t(S_{jt}, \Sigma), \Sigma, v\right]^{Y_{hjt}} \phi(v)\partial v.$$

The corresponding log-likelihood of the data is as follows:

$$(8) \qquad \ell(Y;\Sigma) = \sum_{h=1}^{H} \log\left[L_h(Y_h|A, X;\Sigma)\right].$$

The estimation proceeds by searching for the values of Σ that maximize Equation 8.

The estimation procedure involves iterating over the search for Σ and updating $\lambda(S_{jt}, \Sigma)$ through numerical inversion. Thus, we still use the household-level data to estimate the heterogeneity parameters, Σ. When computing the standard errors for the estimates of Σ, it is necessary to account for heterogeneity entering the utility function both directly as a linear term and indirectly as the term $\lambda(S_{jt}, \Sigma)$. Because we cannot evaluate the high-dimensional integrals in Equation 7 analytically, we use Monte Carlo simulation. We estimate the parameters Σ through simulated maximum likelihood (see Keane 1997).

To recover estimates of the mean taste parameters, $(\alpha, \beta)'$, recall that $\lambda_{jt} = A_j\alpha + X_{jt}\beta + \zeta_{jt}$. Thus, we can run a regression of the functions $\lambda_{jt}(S_{jt}, \Sigma)$ on the product attributes A_j and the marketing variables X_{jt}, where the regression error term is ζ_{jt}. At this stage, we need to be careful about the potential role of price endogeneity. If prices contained in X_{jt} are positively correlated with the error term, ζ_{jt}, ordinary least squares (OLS) estimates of the mean price response coefficient will be biased toward zero. To alleviate this problem, we use an IV procedure.

At this stage, we address the uncertainty in $\lambda(S_{jt}, \Sigma)$ due to the estimation error in Σ by using a parametric bootstrap procedure. For this, we must draw from the asymptotic distribution of our estimates of Σ. For each draw, we compute the corresponding values of λ. Using our bootstrap samples of λ, we can compute the empirical covariance matrix of λ. This matrix is used as a weight in the previous regression. Thus, we use a three-stage least squares estimator that takes into account the uncertainty of λ due to estimation error in Σ and controls for the potential endogeneity of prices using IV.

[2]If we assumed that we have a finite population of H_t distinct consumers, observed aggregate shares would not be deterministic. Instead, conditional on ζ, shares would be distributed multinomial Dirichlet with a mean equal to S_{jt} and a variance of the sampling distribution equal to $S_{jt}(1 - S_{jt})/H_t$. Because this variance is inversely proportional to H_t, it becomes extremely small as H_t becomes large. For this reason, the literature routinely assumes a continuum of consumers to remove the sampling error entirely. Thus, ζ plays the role of the econometric error term in the aggregate model (for a discussion of this issue, see Berry 1994).

[3]We use the contraction mapping that BLP (1995) propose.

[4]Goolsbee and Petrin (2003) use the household sample shares $s_{jt} = \Sigma_h Y_{hjt}/\Sigma_h\Sigma_j Y_{hjt}$, which are imprecise estimates of the aggregate shares, S_{jt}. Because the inversion procedure recovers functions $\lambda(s_{jt}, \Sigma)$, these will also be subjected to the uncertainty surrounding the sample shares. We obtain store shares by aggregating data across all consumers who visit the store, and to the extent that prices and promotions in the category of interest do not influence consumers' store choice decisions, our approach is appropriate.

Discussion of the Estimation Procedure

We now provide a more intuitive discussion of the previously described estimator. The data required for implementation of the proposed approach are the following:

- Household-level data for the sample of households in the market for all store visits made during the time period of interest. Both purchase and no-purchase store visit information is required;
- Store-level data for the same product and geographic market as the household data for all the stores from which the panel households make purchases. These data are the weekly brand sales information for each considered brand. We convert brand sales into shares for each brand and the share of the outside good; and
- Covariate information (e.g., price, features, displays) and attributes for each item.

We now lay out the various steps for the estimation of the model parameters using both household data and store data. The estimation procedure we use is an iterative two-step procedure in which the first step is a maximum likelihood procedure and the second stage involves an IV regression. In general, these steps are interchangeable, and estimation can proceed in the reverse order as well. The following are additional steps for expositional ease:

Step 1: Given starting values for λ_{jt}, for each brand and each time period, we can estimate the parameters of Σ using maximum likelihood and household data. This is standard choice model estimation, but it is even simpler because we no longer need to estimate the mean of the heterogeneity distribution (because it is part of λ_{jt}). We can use either a continuous heterogeneity representation or a latent class, finite mixture approach.

Step 2: Given the estimated values of Σ, we now compute λ_{jt} from the store share data. This step consists of two substeps. First, we use Equation 6 to compute the weekly aggregate shares for each item and the no-purchase alternative conditional on the current values of Σ. Second, we compute the values of λ_{jt} for each brand and time period that equates these aggregated shares to the shares observed in the store data. We can accomplish this computation through the inversion or "contraction mapping" algorithm that BLP (1995) propose.[5]

Step 3: We repeat Steps 1 and 2 until convergence. Convergence occurs when the estimated parameters from two iterations do not differ by more than a value of 1e – 6.

Step 4: After we obtain the converged values for the λ_{jt} "parameters," we recognize that $\lambda_{jt} = A_j\alpha + X_{jt}\beta + \zeta_{jt}$, which is a linear regression with the proviso that the X_{jt} could be correlated with ζ_{jt} because of possible endogeneity. Thus, using appropriate instruments for price, we can estimate the parameters α and β using linear IV procedure.

Before discussing the data and results, we comment on the properties of the proposed methodology. We find that convergence occurs within a few "superiterations." Recall that each such iteration requires convergence of the household model conditional on the mean utilities (which could take several iterations) and the inversion procedure with the store data. As the computational burden on the household data maximum likelihood estimation (MLE) procedure is reduced to the estimation of fewer parameters than usual, each superiteration takes less time than estimating the traditional random coefficients brand choice model. However, because such multiple superiterations are required, the computational time is about five times that of the estimation of a random coefficients model for the data that we describe subsequently.

DATA

We use ACNielsen fabric softeners data from the Denver market collected over a time period of 117 weeks. We use this category for two reasons: First, it is the same one used by Fader and Hardie (1996), who also estimate a SKU-level model with household data without accounting for price endogeneity. Second, because consumers typically purchase only one unit of the product (albeit of different sizes), we do not need to be concerned about the purchase quantity decision. We focus on the top 25 UPCs, which account for 80% of the total sales in the category. Even with 25 UPCs, the unconditional shares of the items in this case are much smaller than Fader and Hardie's because of our inclusion of the no-purchase alternative in the analysis. Thus, no-purchase serves as the 26th item in the analysis. Because the unobserved item characteristics in our previous discussion were attributed largely to retail environment factors, we restricted our attention to households that made purchases only at the single largest chain to avoid computing chain-specific characteristics. The household data contain information on all shopping trips the households made, including those on which no softener was purchased. The store data contain the unit sales, shelf prices, and whether a brand was featured or displayed only or both featured and displayed. Descriptive statistics for the data appear in Tables 1 and 2.

At the store level, we first calculate conditional shares from total sales. We then convert these into market shares using the observed household sample no-purchase shares. Implicitly, we assume that the total market size, H_t, is weekly store traffic.[6] In this formulation, the total market size is exogenous. In other words, fabric softener prices do not influence store choice, and thus the total pool of potential buyers is fixed. Nevertheless, by modeling the no-purchase probability, we allow the category size to expand and contract over time as a function of prices. These assumptions are standard in the choice-modeling literature.[7]

A total of 657 households satisfied the criterion of making a purchase among the 25 UPCs and shopping at the largest chain. Table 1 indicates that approximately 5% of all these households' trips result in a purchase in the category. Table 2 indicates considerable variation in the (unconditional) shares across the UPCs; the largest shares are

[5]When the household panel data are not sparse and aggregation of shares to the weekly level provides reasonable share estimates, we can simply replace the store share data with aggregated household data. Note that we need to be careful to correct the standard errors for the sampling error in these shares (Goolsbee and Petrin 2003).

[6]Alternatively, we could also use the total customer count of the store or the weekly store traffic, which constitutes the size of the potential market, to calculate the aggregate no-purchase shares (Chintagunta 2002).

[7]Slade (1995) surveyed retail managers of the stores in a comparable scanner database. The managers responded that, on average, fewer than 10% of total shoppers in a given week shop across stores. Moreover, the belief is that store choice is not dependent on item prices but rather on a price index representing the overall price level in a store.

Table 1
DESCRIPTIVE STATISTICS FOR THE DATA

Variable	*Value*
Total households	657
Total trips	31,544
Purchase trips	1488
Trips/household	48
Weeks	82

approximately 4–5 times larger than the smallest ones. Notably, across the 82 weeks, it appears that the average shares of the UPCs are roughly comparable at the store and household level. There is also considerable variation in retail prices (operationalized as price per use) across the UPCs. Table 2 also has a column for wholesale prices. This is the weekly price that each manufacturer charges to the retailers in the Denver market. We include these wholesale prices in the analysis to act as instruments for retail prices (for a discussion, see Chintagunta 2002). The assumption here is that the wholesale prices are correlated with the retail prices but are unlikely to be correlated with in-store coupons and other retail-level factors that are reflected in the unobserved characteristics term. Thus, wholesale prices are reasonable instruments for retail prices. The differences between the retail and wholesale prices in Table 2 indicate the retail markups in this chain. The average retail margin ([retail price – wholesale price]/retail price) is approximately 18% in this category. Notably, we find that the retailer prices two Snuggle UPCs (numbers 2 and 3) just below wholesale price. The big margin UPCs are 7, 11, 17, and 19, which tend to be the smaller-sized UPCs. The promotion variables in Table 2 represent the proportion of weeks that each brand is displayed. Finally, Table 2 provides the attribute levels for each of the four attributes that characterize each UPC. The four attributes are the brand (Bounce: 1, Cling Free: 2, Downy: 3, Final Touch: 4, Snuggle: 5, and Toss 'n Soft: 6), form (sheets: 1, and liquid: 2); scent (regular: 1, unscented: 2, sunrise fresh: 3, morning fresh: 4, cuddle-up fresh: 5, and ultra blue: 6); and size (small: 1, medium: 2, large: 3, and extra large: 4).

The results we present here are for a continuous heterogeneity distribution with a joint distribution on the brand intercepts and price sensitivity parameter. All other parameters for the attribute effects and promotional effects have independent normal distributions. Such an approach results in a flexible substitution pattern across UPCs without needing to estimate too many parameters. We perform the estimation using simulated maximum likelihood. We use 30 draws and check the sensitivity of our results to this number.[8]

EMPIRICAL RESULTS

Simulation Results

Before fitting the model to our household and store data, we ran the procedure using simulated data. In particular, we created a synthetic data set that consisted of weekly population market shares along with a corresponding random subsample of households' choice histories. In addition, we included an error term (i.e., the unmeasured brand characteristics) that was correlated with prices. We find that the

[8]We used 10, 20, 30, 40, and 50 draws. Estimates showed only minor changes from 30 draws onward. Thus, we chose this as the most parsimonious representation. The results from the finite mixture approach are available on request.

Table 2
UPC-LEVEL DESCRIPTIVE STATISTICS

UPC Number	*Brand*	*Form*	*Scent*	*Size*	*Share Household*	*Share Store*	*Retail Price Per Use*	*Wholesale Price Per Use*	*Display*
1	Bounce	Sheet	Regular	Medium	.0047	.0048	5.79	4.74	.13
2	Snuggle	Liquid	Cuddle-up fresh	Small	.0032	.0025	10.26	10.72	.09
3	Snuggle	Liquid	Morning fresh	Small	.0026	.0017	10.48	10.72	.11
4	Bounce	Sheet	Unscented	Medium	.0031	.0036	5.81	4.74	.02
5	Snuggle	Sheet	Regular	Medium	.0030	.0027	5.35	4.71	.09
6	Bounce	Sheet	Regular	Extra large	.0021	.0023	5.03	4.39	.03
7	Downy	Liquid	Regular	Large	.0021	.0036	5.92	4.30	.00
8	Downy	Liquid	Ultra blue	Small	.0019	.0021	13.11	12.35	.00
9	Snuggle	Sheet	Morning fresh	Medium	.0021	.0016	5.34	4.71	.09
10	Snuggle	Sheet	Regular	Large	.0016	.0014	5.43	4.49	.04
11	Toss 'n Soft	Sheet	Regular	Small	.0012	.0013	5.78	3.54	.00
12	Snuggle	Liquid	Morning fresh	Medium	.0016	.0010	11.03	9.50	.06
13	Snuggle	Liquid	Cuddle-up fresh	Medium	.0015	.0014	11.01	9.50	.05
14	Downy	Sheet	Regular	Small	.0018	.0022	7.65	5.90	.00
15	Downy	Liquid	Ultra blue	Medium	.0017	.0019	11.52	9.65	.00
16	Snuggle	Sheet	Regular	Small	.0014	.0016	7.01	5.81	.00
17	Bounce	Sheet	Regular	Small	.0016	.0030	7.95	4.74	.00
18	Downy	Sheet	Regular	Medium	.0013	.0018	5.91	4.76	.00
19	Cling Free	Sheet	Regular	Medium	.0010	.0012	5.26	3.89	.02
20	Final Touch	Liquid	Regular	Small	.0010	.0008	8.82	7.19	.00
21	Snuggle	Sheet	Morning fresh	Small	.0012	.0009	7.02	5.81	.00
22	Snuggle	Sheet	Morning fresh	Large	.0011	.0007	5.46	4.49	.05
23	Bounce	Sheet	Regular	Large	.0016	.0014	5.95	4.56	.00
24	Final Touch	Sheet	Regular	Small	.0015	.0010	5.15	3.74	.00
25	Downy	Sheet	Sunrise fresh	Medium	.0011	.0008	5.91	4.76	.00

Notes: Shares are unconditional on purchase, prices are in dollars/item, and display refers to the proportion of times items are displayed.

proposed procedure recovers parameter estimates that are statistically unbiased. That is, we fail to reject the hypothesis that our parameter estimates are equal to the true values. In contrast, we find significant biases when we fit the household data using only usual maximum likelihood methods.

Model Fit and Predictive Ability

Before presenting the parameter estimates, we discuss the fit of the proposed model and its predictive ability. As a benchmark, we use the results from the model that uses the individual household data and the MLE method. This is the random coefficients choice model but without the unobserved attribute. In terms of model fit, the value of the Bayesian information criterion for the proposed model is 19,530, and that of the benchmark model is 19,873. This implies a better in-sample fit for the proposed model. We then reestimated the model parameters after creating a holdout sample of 78 households, and we used these parameters to forecast the shares of the different brands across time periods for these 78 households. In performing this analysis, we did not use information on the unobserved attribute in forecasting under the proposed model, because the unobserved attribute contains information from the store data, which are the aggregation of data from all households including the holdout households. We computed the mean absolute percentage deviations (MAPD) between the predicted and actual shares averaged across all time periods. Then, we computed a share weighted average of this MAPD measure across all SKUs. The benchmark model yielded an MAPD measure of .51; for the proposed model, the value was .33. Thus, the proposed model outperforms the benchmark model on the forecasting criterion. A caveat to this analysis is that the households in the holdout sample are not systematically different from the estimation households.

Results from Scanner Data

Table 3 provides the results from the three models. The first model, Individual Data–MLE (ID–MLE), is the random coefficients choice model using household-level data but without the unobserved attribute. This is the continuous heterogeneity analog to Fader and Hardie's (1996) approach. The second model, Combo Data–OLS (CD–

Table 3

MODELS WITH HETEROGENEITY

		ID–MLE		*CD–OLS*		*CD–IV*	
Attribute	*Variable*	*Parameter*	*Standard Error*	*Parameter*	*Standard Error*	*Parameter*	*Standard Error*
Mean Effects							
Brand	Bounce	1.34	.17	2.25	.09	2.73	.10
	Cling Free	.04	.23	.55	.10	.84	.11
	Downy	−1.24	.28	.71	.08	1.10	.09
	Final Touch	−1.04	.12	−.96	.08	−.76	.09
	Snuggle	.94	.13	.87	.09	1.23	.10
	Toss 'n Soft	.00		.00		.00	
Form	Sheets	−2.11	.08	−2.61	.05	−2.84	.06
	Liquid	.000		.00		.00	
Scent	Regular	−3.05	.12	−2.26	.07	−2.32	.07
	Unscented	−3.39	.21	−2.25	.11	−2.46	.12
	Sunrise fresh	−3.38	.38	−3.73	.10	−3.84	.10
	Morning Fresh	−3.25	.14	−2.84	.08	−2.86	.08
	Cuddle-up fresh	−3.10	.16	−2.69	.09	−2.67	.09
	Ultra blue	.00		.00		.00	
Size	Small	1.52	.13	1.77	.09	2.30	.11
	Medium	.96	.10	1.08	.09	1.50	.10
	Large	−.10	.11	−.32	.08	−.03	.09
	Extra large	.00		.00		.00	
Covariates	Price	−.65	.02	−.78	.01	−.86	.01
	Display	−.09	.15	.17	.12	.00	.13
Standard Deviation of Heterogeneity Distribution							
Brand	Bounce[a]	2.19	.16	1.96	.13	1.96	.13
	Cling Free[a]	.95	.39	.84	.29	.84	.29
	Downy[a]	2.27	.18	1.48	.11	1.48	.11
	Final Touch[a]	2.68	.21	2.48	.17	2.48	.17
	Snuggle[a]	1.89	.09	1.81	.08	1.81	.08
Form	Sheets	.60	.06	.42	.06	.42	.06
Scent	Regular	.96	.06	.85	.05	.85	.05
	Unscented	1.14	.13	.41	.08	.41	.08
	Sunrise fresh	2.15	.36	1.93	.11	1.93	.11
	Morning fresh	1.21	.11	1.62	.08	1.62	.08
	Cuddle-up fresh	.13	.13	.53	.07	.53	.07
Size	Small	1.29	.08	1.15	.07	1.15	.07
	Medium	.87	.08	.99	.07	.99	.07
	Large	.01	.09	.01	.08	.01	.08
Covariates	Price[a]	.15	.03	.18	.02	.18	.02
	Display	1.38	.16	.59	.10	.59	.10

[a]These parameters refer to the diagonal of the covariance matrix of the joint distribution of preferences and price sensitivities.

OLS), combines information from household and store levels but runs an OLS regression rather than an IV regression to recover the mean parameters from the store data. The third specification, Combo Data–IV (CD–IV), is the same as the second model except that endogeneity is accounted for with the IV procedure.

In examining the parameters that characterize the mean effects of the attributes and the marketing variables, we find the following: The ranges of the attributes' effects are greater under the CD–IV specification than under the ID–MLE specification for all four attributes. Furthermore, the intrinsic preference for Downy (a Procter & Gamble [P&G] brand) is the lowest among all the brands under the household data–only specification. Under the IV specification, only Bounce (another P&G product) and Snuggle have higher intrinsic preferences. Given P&G's large investments in Downy advertising and innovation (e.g., the Downy ball), the results from the CD–IV model appear to be more plausible. One of the key substantive implications that Fader and Hardie (1996) note is related to using the results for forecasting the sales of new SKUs. Our results for the Downy brand indicate that these forecasts are influenced by the model that is chosen. To the extent that the proposed specification accounts for both endogeneity and heterogeneity issues, forecasts from such a specification may be more acceptable. We revisit the forecasting issue subsequently.

Comparison of the results in Table 3 with a model that does not account for heterogeneity (we do not report this here) indicates that the mean price effects from the models in Table 3 (that account for heterogeneity) are approximately 10% larger in magnitude. This is consistent with the elasticity results we report subsequently. At the same time, we find that there is evidence of an endogeneity bias in these data. In particular, the price coefficient in the CD–OLS model is –.78, and the price coefficient in the CD–IV model is –.86; the difference is statistically significant. Thus, for these data, we find evidence for both heterogeneity and an endogeneity bias. Note that we are able to establish this only because we combine the two data sources. With access only to household-level data, we would not have had enough information to uncover the effects of the unobserved attribute. In this situation, combining the two data sources at the household and store levels enables us to account for heterogeneity and price endogeneity. Notably, when we compare the precision of the parameters for the ID–MLE with the CD–OLS approach (Columns 4 and 6 in Table 3), we find the latter to be more precise in most cases.

In comparing the attribute effects across the OLS and IV approaches with the CD, we find that, as we expected, the IV procedure raises the standard errors of the parameter estimates. However, this reduction in efficiency is not as large as the improvement obtained from combining the two data sources. Thus, although in general correcting for the endogeneity bias should be traded off against the reduced efficiency that is inherent in the IV procedure, in our case, we are not severely affected by this issue.

Turning to the estimates of the heterogeneity distribution, we estimate a joint distribution for brand preferences and price sensitivities and assume that the preferences for the other attributes are distributed univariate normal. The results in Table 3 indicate that there is considerable heterogeneity in attribute preferences across households. In particular, households appear to vary in their preferences for brands, form, scent, and size. Although this finding is consistent across the model specifications, note that of the 14 parameters we allow to be heterogeneous across households, the ID–MLE model implies a higher level of heterogeneity in 10 of the cases. In addition, we find that there is significant heterogeneity in the price and display sensitivities across households and that the amount of heterogeneity in price sensitivities is comparable across model specifications. An important implication of these heterogeneity findings on attributes and price effects is that not accounting for the unobserved attribute (i.e., the ζ_{jt} in Equation 2) can potentially overstate the amount of heterogeneity across households because it implies more heterogeneity in 11 of 16 cases. This finding has implications for activities such as targeting that exploit the extent of heterogeneity across households.

Next, we discuss the own price elasticities from the three models. To observe how the different specifications influence the demand elasticities for the individual brands, we present these elasticities in Table 4. Table 4 indicates that in magnitude terms, the elasticities are ordered as follows: CD–IV > CD–OLS > ID–MLE. Consistent with the endogeneity bias effect, we find that the elasticities from the CD–IV model are the highest of the three. Notably, we find that the CD–OLS procedure also yields slightly larger own elasticities than the household data MLE model. For these data, this implies that accounting for the presence of the unobserved attributes without considering their correlation with prices increases own price elasticities somewhat.

Based on the data in Table 4, if the objective is to obtain the correct price elasticities for a market and if the mean effects of price are better represented by store data and the heterogeneity distribution by the household data, we recommend combining the information from the two sources. Furthermore, while carrying out the joint estimation, it is important to account for price endogeneity in the data properly. Failure to account for any of these three features (i.e., combining the data when household data are sparse, accounting for heterogeneity, and accounting for price endogeneity) may lead to incorrect estimates of price elasticities. In contrast, our proposed approach provides robust estimates of the model parameters and thus the marginal effects of marketing variables such as price elasticities.

Substantive Implications

Thus far, our results indicate four areas in which the substantive implications from the data differ under our proposed approach compared with the extant household data–only approach that is prevalent in the literature: (1) The estimated mean effects of attributes differ across the two approaches, and the ranges of the attribute effects estimated are greater under the proposed approach that combines data sources. (2) Consumers are more price elastic than a model that uses only household data implies. Note that using only household data precludes our ability to account for endogeneity in our application because of sparse data. (3) Not accounting for the unobserved attribute could overstate the extent of heterogeneity in the data. This has implications for tactics such as targeting, one of the reasons marketers might be interested in using household-level data. (4) An implication that follows from these three points is that sales fore-

Table 4

OWN PRICE ELASTICITIES HETEROGENEOUS MODELS

UPC Number	Brand	Form	Scent	Size	ID–MLE Mean	ID–MLE Standard Deviation	CD–OLS Mean	CD–OLS Standard Deviation	CD–IV Mean	CD–IV Standard Deviation
1	Bounce	Sheet	Regular	Medium	–4.21	.38	–4.74	.39	–5.17	.44
2	Snuggle	Liquid	Cuddle-up fresh	Small	–5.53	.82	–6.24	.88	–6.97	1.01
3	Snuggle	Liquid	Morning fresh	Small	–5.97	.99	–7.12	1.15	–7.85	1.29
4	Bounce	Sheet	Unscented	Medium	–4.19	.41	–4.80	.42	–5.23	.47
5	Snuggle	Sheet	Regular	Medium	–3.21	.31	–3.78	.32	–4.18	.37
6	Bounce	Sheet	Regular	Extra large	–3.69	.57	–4.04	.57	–4.41	.63
7	Downy	Liquid	Regular	Large	–3.99	.00	–4.41	.00	–4.82	.01
8	Downy	Liquid	Ultra blue	Small	–7.56	.15	–7.59	.13	–8.54	.16
9	Snuggle	Sheet	Morning fresh	Medium	–3.49	.31	–4.19	.35	–4.59	.39
10	Snuggle	Sheet	Regular	Large	–3.38	.30	–3.92	.33	–4.34	.37
11	Toss 'n Soft	Sheet	Regular	Small	–2.94	.10	–3.39	.11	–3.83	.13
12	Snuggle	Liquid	Morning fresh	Medium	–6.16	.81	–7.27	.93	–8.06	1.06
13	Snuggle	Liquid	Cuddle-up fresh	Medium	–5.81	.91	–6.90	1.04	–7.69	1.19
14	Downy	Sheet	Regular	Small	–5.38	.00	–5.26	.00	–5.82	.00
15	Downy	Liquid	Ultra blue	Medium	–6.61	.58	–7.39	.66	–8.21	.77
16	Snuggle	Sheet	Regular	Small	–4.13	.18	–4.72	.19	–5.24	.22
17	Bounce	Sheet	Regular	Small	–5.53	.00	–5.55	.01	–6.13	.01
18	Downy	Sheet	Regular	Medium	–3.85	.11	–4.34	.12	–4.78	.13
19	Cling Free	Sheet	Regular	Medium	–3.17	.10	–3.75	.11	–4.14	.13
20	Final Touch	Liquid	Regular	Small	–5.71	.44	–6.60	.45	–7.25	.52
21	Snuggle	Sheet	Morning fresh	Small	–4.40	.23	–5.22	.27	–5.73	.30
22	Snuggle	Sheet	Morning fresh	Large	–3.61	.36	–4.32	.42	–4.73	.46
23	Bounce	Sheet	Regular	Large	–4.38	.00	–4.71	.00	–5.16	.00
24	Final Touch	Sheet	Regular	Small	–3.73	.23	–4.46	.25	–4.84	.28
25	Downy	Sheet	Sunrise fresh	Medium	–3.47	.11	–4.13	.12	–4.57	.14

casts for new SKUs are likely to differ under the two approaches.

Here, we focus on the forecasting issue, one of the main substantive implications that Fader and Hardie (1996) provide using the attributes-based approach. Unlike their scenario, however, we do not have any actual line extensions. Thus, we contrasted the forecasts obtained from the two methods (i.e., ID–MLE and CD–IV) for hypothetical line extensions. We chose one large brand (Downy) and one small brand (Cling Free) for this purpose. We "launched" a sheet form of the former brand and a liquid version of the latter brand. The Downy sheets had the morning fresh scent and were available in a medium size (this corresponds to a product sold by rival, Snuggle), and Cling Free liquid had a regular scent and was available in a large size (this corresponds to a successful Downy product). We introduced these products one at a time and computed the predicted shares. For simplicity, we computed the shares at the levels of marketing activities that corresponded to the imitated products and in the absence of unobserved attributes for both products.[9]

We find that the ID–MLE model predicts unconditional shares of .053% and .365% for the Downy and Cling Free line extensions, respectively. The corresponding numbers under the CD–IV model are .102% and .401%. Note that these numbers are small because of the large share of the no-purchase option (approximately 94%). These share predictions indicate that for the Downy line extension, the proposed model forecasts a share that is almost twice as large as that of the traditional approach. However, for the Cling Free product, the forecasts are closer together. The differences in results across the two products stem from the differences in the estimated attribute effects for Downy and Cling Free across specifications. In particular, recall that the estimate for the Downy attribute level was different across models. However, the Cling Free effect was much closer. An examination of the predicted shares conditional on purchasing in the category suggests the following predictions: .97% and 7.5% under the ID–MLE model and 1.95% and 7.85% under the CD–IV model for Downy and Cling Free, respectively. These numbers reflect the differences in the unconditional shares.

In Table 5, we provide a source of share analysis for the conditional (on purchase) shares for the two line extensions. Table 5 reveals large differences across models with respect to where the share for the line extension is coming from. For the Downy line extension, we find that the ID–MLE model predicts the top three sources as being other Downy SKUs (7, 8, and 15), whereas the CD–IV model predicts share stealing from one Downy SKU (7), from the SKU that the extension imitates (Snuggle, 9), and from another Snuggle SKU with the same scent as the line extension. Therefore, the Downy brand manager contemplating the launch of the line extension may be less willing to do so under the ID–MLE model, which predicts cannibalization, than under the CD–IV model, which predicts cannibalization and share stealing from a large rival. In the case of the Cling Free line extension, the two models are consistent in two of the top three SKUs from which switching is predicted. However, for the third, the CD–IV model again predicts the more plausible scenario of share stealing from the SKU that the line extension imitates (Downy, 7). Thus, our source of share analysis implies that there are important differences across models that might have substantive and managerial consequences.

[9]For the Downy (Cling Free) line extension, we used the prices of the rival Snuggle (Downy) product to perform the forecast. In addition, we assumed that the values of the unobserved attribute were zero for these line extensions.

Table 5

SOURCE-OF-CONDITIONAL SHARE ANALYSIS FOR NEW SKUS

					Source of Share		*Source of Share*	
					Downy	*Extension*	*Cling Free*	*Extension*
UPC Number	*Brand*	*Form*	*Scent*	*Size*	*ID–MLE*	*CD–IV*	*ID–MLE*	*CD–IV*
1	Bounce	Sheet	Regular	Medium	–.05	–.11	–.35	–.60
2	Snuggle	Liquid	Cuddle-up fresh	Small	–.03	–.06	–.58	–.65
3	Snuggle	Liquid	Morning fresh	Small	–.05	–.21	–.24	–.26
4	Bounce	Sheet	Unscent	Medium	–.03	–.12	–.18	–.34
5	Snuggle	Sheet	Regular	Medium	–.04	–.06	–.39	–.34
6	Bounce	Sheet	Regular	Extra large	–.06	–.10	–.41	–.58
7	Downy	Liquid	Regular	Large	–.11	–.26	–.26	–.78
8	Downy	Liquid	Ultra blue	Small	–.08	–.08	–.16	–.24
9	Snuggle	Sheet	Morning fresh	Medium	–.07	–.16	–.29	–.22
10	Snuggle	Sheet	Regular	Large	–.01	–.02	–.29	–.16
11	Toss 'n Soft	Sheet	Regular	Small	–.03	–.03	–.79	–.57
12	Snuggle	Liquid	Morning fresh	Medium	–.01	–.03	–.22	–.12
13	Snuggle	Liquid	Cuddle-up fresh	Medium	–.01	–.01	–.24	–.13
14	Downy	Sheet	Regular	Small	–.04	–.05	–.05	–.09
15	Downy	Liquid	Ultra blue	Medium	–.08	–.08	–.19	–.23
16	Snuggle	Sheet	Regular	Small	–.02	–.03	–.56	–.45
17	Bounce	Sheet	Regular	Small	–.02	–.05	–.20	–.40
18	Downy	Sheet	Regular	Medium	–.03	–.04	–.05	–.08
19	Cling Free	Sheet	Regular	Medium	–.03	–.04	–1.19	–.87
20	Final Touch	Liquid	Regular	Small	–.01	–.03	–.15	–.11
21	Snuggle	Sheet	Morning fresh	Small	–.04	–.13	–.12	–.12
22	Snuggle	Sheet	Morning fresh	Large	–.03	–.05	–.28	–.15
23	Bounce	Sheet	Regular	Large	–.02	–.04	–.20	–.23
24	Final Touch	Sheet	Regular	Small	–.03	–.07	–.10	–.08
25	Downy	Sheet	Sunrise fresh	Medium	–.05	–.07	–.02	–.02
26[a]	Downy	Sheet	Morning fresh	Medium	.97	1.95		
27[a]	Cling Free	Liquid	Regular	Large			7.50	7.85

[a]Numbers for line extensions (26 and 27) are predicted shares in percentage terms after the outside good is excluded.

Notes: Source-of-share numbers correspond to reduction in shares from the existing SKUs that make up the shares of the line extensions. Note that column numbers sum to zero because we are examining conditional shares.

Incorporating Purchase Quantities

A limitation of our model formulation is that it is not suitable for categories in which consumers make quantity decisions in addition to purchase incidence and brand choice decisions. To address this issue, we use the random coefficients analog of the discrete continuous model that Chiang (1992) uses. We augment the model to account for potential price endogeneity related to unobserved (to the researcher) brand characteristics (see also Nair, Dubé, and Chintagunta 2005).[10] We calibrate the incidence–choice–quantity model using refrigerated orange juice purchases of 429 households in the same Denver market we used for the fabric softeners. Consumers chose among four main alternatives that accounted for a majority of purchases: Minute Maid, 64 ounces; Tropicana, 64 ounces; Minute Maid, 96 ounces; and Tropicana, 96 ounces. Approximately 8% of the weeks had no household purchases of one or more of the four alternatives, which required us to use our proposed methodology to account for both endogeneity and heterogeneity. Consistent with our results in the fabric softeners category, we find that the price effect is biased toward zero if endogeneity is not accounted for. The price coefficient from the CD–IV model is 33% greater in magnitude than that of the ID–MLE model. In terms of the elasticities, we computed purchase incidence elasticities, conditional and unconditional brand choice elasticities, and purchase quantity elasticities. We find that all our elasticity estimates lie within the range that Bell, Chiang, and Padmanabhan (1999) report. In examining the differences across model specifications of these elasticities, we find the following: The purchase incidence elasticities for all four brands are greater in the CD–IV case than in the ID–MLE case. The relative ordering of these elasticities across the brands differs under the two models. Whereas the proposed model implies greater elasticities for the two 64-ounce products, the model that does not account for endogeneity indicates that Tropicana, 64 ounces, has the highest elasticity, followed by Minute Maid, 96 ounces; Minute Maid, 64 ounces; and Tropicana, 96 ounces. Thus, we find a shift in the relative purchase incidence elasticities in addition to the overall levels of the elasticities when endogeneity is not accounted for. Differences in relative elasticities are also observed for the conditional brand choice decision. Finally, from the purchase quantity elasticities, we find that these elasticities are small for the orange juice category. However, differences in the relative ordering of brands across model specifications continue to persist in this case. The small magnitudes for quantity elasticities are not surprising in this case, because we have already accounted for size elasticities through our operationalization of the choice alternatives.

[10]Because of space considerations, we do not present the details of the model and its derivation along with the empirical results. They are in an Appendix that is available on request.

As with the fabric softeners category, we can derive substantive implications from our results in the orange juice category. In particular, we note the recent literature on decomposing the sales "bump" due to promotions (Van Heerde, Gupta, and Wittink 2003; Van Heerde, Leeflang, and Wittink 2005). Specifically, these studies take the price elasticity of demand and decompose it into the corresponding promotional impacts on unit sales. Our analysis indicates that the nature of these effects depends on the ability to account properly for price endogeneity and consumer heterogeneity. To the extent that such controls are possible only when we combine both household and store data, the method we propose herein makes a contribution to the literature on sales decompositions.

CONCLUSIONS

The marketing literature has established the importance of accounting for heterogeneity and endogeneity with scanner data. However, there might be several situations in which it is not possible to account for these issues using only household panel data (e.g., Villas-Boas and Zhao 2005) or store sales data (e.g., Besanko, Gupta, and Jain 1998) because of (1) the sparseness of panelist purchases with the household data when the analysis must be carried out at the SKU level (e.g., Fader and Hardie 1996) and (2) the inability to account for heterogeneity and purchase quantities with store data. For such situations, we propose a simple approach that combines the information from household scanner-panel data with that from store data. In particular, we attempt to exploit each source of information for that which it is most useful: heterogeneity in the case of panel data and the mean effects of marketing activities in the case of store data. Thus, we are able to obtain a single set of estimates that account for heterogeneity across households and endogeneity of prices. Our approach is directly relevant to retailers that routinely track their aggregate sales and the purchases of loyalty card–holding customers.[11]

Our empirical results from the fabric softener category at the SKU level provide several substantive implications. The estimated mean effects of attributes are affected when endogeneity is not accounted for; the ranges of the attribute effects estimated are larger under the proposed approach than under an approach in which endogeneity is ignored. Not accounting for endogeneity appears to bias the estimated price effect and the corresponding elasticities toward zero. More significantly, ignoring the unobserved attribute could overstate the extent of heterogeneity in the data. Because of differences in parameter estimates, we also find that sales forecasts for new SKUs differ under the proposed approach compared with an approach that uses only household data without accounting for price endogeneity.

The current approach assumes that consumer choices are static in nature. However, a growing body of literature has documented several important dynamic elements in consumer choices. For example, myopic consumers may be influenced by habit persistence in brand choice (Keane 1997) and by the accumulation of inventories (Ailawadi and Neslin 1998); thus, historic choice behavior could influence current demand. Fully rational consumers might anticipate these dynamics and, in response, make forward-looking decisions. For example, forward-looking consumers would need to solve a complex dynamic program if they anticipate habit persistence (Erdem 1996) or if price expectations influence the timing of stockpiling (Erdem, Imai, and Keane 2003).[12] An implication of ignoring such dynamics is that our current analysis can potentially understate the magnitude of price elasticities (e.g., if households with higher inventory levels are less likely to respond to a price cut). However, this limitation would apply to all the approaches used in this study. Further research would benefit from resolving formally how to reconcile these dynamics with the corresponding aggregate demand system. Although it is unlikely that store data alone could reveal individual dynamics, combining such data with household data might be a worthwhile starting point.

[11]The proposed approach may have limitations for researchers or for data-gathering services, such as Information Resources Inc. or ACNielsen, because some retailers, such as Wal-Mart, do not make their aggregate sales data available.

[12]Related literature has used lead and lag prices to proxy for these dynamic effects in consumer demand. Experimentation with such variables in the context of our data led to very small effect sizes. Moreover, the inclusion of lead price effects raises another potential endogeneity problem that is beyond the scope of this article. Thus, we do not include these proxies in the current analysis.

REFERENCES

Ailawadi, Kusum L. and Scott A. Neslin (1998), "The Effect of Promotion on Consumption: Buying More and Consuming It Faster," *Journal of Marketing Research*, 35 (August), 390–98.

Allenby, Greg and Peter Rossi (1999), "Marketing Models of Consumer Heterogeneity," *Journal of Econometrics*, 89 (1–2), 57–78.

Bell, D.R., J. Chiang, and V. Padmanabhan (1999), "The Decomposition of Promotional Response: An Empirical Generalization," *Marketing Science*, 18 (4), 504–526.

Berry, S.T. (1994), "Estimating Discrete-Choice Models of Product Differentiation," *Rand Journal of Economics*, 25 (2), 242–62.

———, J. Levinsohn, and A. Pakes (1995), "Automobile Prices in Market Equilibrium," *Econometrica*, 60 (4), 841–90.

Besanko D., J.P. Dubé, and S. Gupta (2003), "Competitive Price Discrimination Strategies in a Vertical Channel Using Aggregate Retail Data," *Management Science*, 49 (9), 1121–38.

———, S. Gupta, and D. Jain (1998), "Logit Demand Estimation Under Competitive Pricing Behavior: An Equilibrium Framework," *Management Science*, 44 (11), 1533–47.

Bodapati, Anand and Sachin Gupta (2004), "The Recoverability of Segmentation Structure from Store-Level Aggregate Data," *Journal of Marketing Research*, 41 (August), 351–64.

Bronnenberg, Bart J., Peter E. Rossi, and Naufel J. Vilcassim (2005), "Structural Modeling and Policy Simulation," *Journal of Marketing Research*, 42 (February), 22–29.

Chiang, J. (1992), "A Simultaneous Approach to the Whether, What, and How Much to Buy Questions," *Marketing Science*, 10 (4), 297–315.

Chintagunta, Pradeep K. (2002), "Investigating Category Pricing Behavior in a Retail Chain," *Journal of Marketing Research*, 39 (May), 141–54.

———, J.P. Dubé, and K.Y. Goh (2005), "Beyond the Endogeneity Bias: The Effect of Unmeasured Brand Characteristics on Household-Level Brand Choice Models," *Management Science*, forthcoming.

———, Dipak C. Jain, and Naufel J. Vilcassim (1991), "Investigating Heterogeneity in Brand Preferences in Logit Models for Panel Data," *Journal of Marketing Research*, 28 (November), 417–28.

Erdem, T. (1996), "A Dynamic Analysis of Market Structure Based on Panel Data," *Marketing Science*, 15 (4), 359–78.

———, S. Imai, and M. Keane (2003), "A Model of Consumer Brand and Quantity Choice Dynamics Under Price Uncertainty," *Quantitative Marketing and Economics*, 1 (1), 5–64.

Fader, Peter S. and Bruce G.S. Hardie (1996), "Modeling Consumer Choice Among SKUs," *Journal of Marketing Research*, 33 (November), 442–52.

Goolsbee, A. and A. Petrin (2003), "The Consumer Gains from Direct Broadcast Satellites and the Competition with Cable TV," *Econometrica*, 72 (2), 351–81.

Gupta, Sachin, Pradeep K. Chintagunta, Anil Kaul, and Dick R. Wittink (1996), "Do Household Scanner Panels Provide Representative Inferences from Brand Choices: A Comparison with Store Data," *Journal of Marketing Research*, 33 (November), 383–98.

Kamakura, Wagner A. and Gary J. Russell (1989), "A Probabilistic Choice Model for Market Segmentation and Elasticity Structure," *Journal of Marketing Research*, 26 (November), 379–90.

Keane, M.P. (1997), "Modeling Heterogeneity and State Dependence in Consumer Choice Behavior," *Journal of Business and Economic Statistics*, 15 (3), 310–27.

Nair, H., J.P. Dubé, and P.K. Chintagunta (2005), "Accounting for Primary and Secondary Demand Effects with Aggregate Data," *Marketing Science*, forthcoming.

Slade, M. (1995), "Product Rivalry and Multiple Strategic Weapons: An Analysis of Price and Advertising Competition," *Journal of Economics and Management Strategy*, 4 (3), 445.

Van Heerde, Harald J., Sachin Gupta, and Dick R. Wittink (2003), "Is 75% of the Sales Promotion Bump Due to Brand Switching? No, Only 33% Is," *Journal of Marketing Research*, 40 (November), 481–91.

———, P.S.H. Leeflang, and D.R. Wittink (2005), "Decomposing the Sales Promotion Bump with Store Data," *Marketing Science*, 23 (3), 317–34.

Villas-Boas, J. Miguel and Ying Zhao (2005), "Retailer, Manufacturers, and Individual Consumers: Modeling the Supply Side in the Ketchup Marketplace," *Journal of Marketing Research*, 42 (February), 83–95.

Yang, S., Y. Chen, and G. Allenby (2003), "Bayesian Analysis of Simultaneous Demand and Supply," *Quantitative Marketing and Economics*, 1 (3), 251–75.