ANAND V. BODAPATI and SACHIN GUPTA*

The authors focus on estimating a latent-class choice model with consumer response segments when only store-level aggregate data are available. Most of the proposed methodologies in the marketing literature require household panel data, which can be difficult to obtain. There is a growing stream of work in marketing and in empirical industrial organization that estimates segmentation structure with aggregate data. This article is a careful attempt to understand the extent to which disaggregate structure in the form of a latent-class model can be recovered from aggregate data. The authors show that under specific assumptions and when the household-level model is correctly specified, most of a latent-class segmentation structure is *identifiable* even if only store-level aggregate data are available. Therefore, the store data–based estimates for the latent-class model are *consistent*. In other words, the mean absolute deviation (MAD) of the estimates goes to zero with infinite sample size. To assess how well latent-class structure can be estimated from store data sets of sample sizes that are comparable to those in real life, the authors simulate more than 60,000 store data sets and compute the consequent model estimates and their estimation errors. The results show that the MAD of the latent-class estimates diminishes much more slowly with store data than with household data. Moreover, the rate is so slow that obtaining estimates with reasonably small MADs often requires unreasonably large sample sizes. The authors' simulations offer guidance on conditions that favor obtaining more accurate estimates from store-level data.

The Recoverability of Segmentation Structure from Store-Level Aggregate Data

Manufacturers of packaged consumer goods rely heavily on two types of Universal Product Code scanner data: point-of-sale data and household panel data. The two data sources are typically viewed as complementary; that is, each has its unique set of uses and applications. Point-of-sale data are used primarily to monitor category and brand performance in terms of volume and market share and to measure the effectiveness of firms' marketing activities, such as trade and consumer promotions. Household panel data are used to track new product trial and repeat rates, to understand demographics of buyers and nonbuyers, to examine brand switching and loyalty patterns, and to understand differences between consumers' buying characteristics for purposes of segmentation and targeting. Because of the complementary uses of the data, most large consumer packaged goods manufacturers in the United States purchase both types of data from ACNielsen and/or Information Resources Inc., the two major vendors of syndicated scanner data (Prevision Corporation 1995).

Point-of-sale sales data (or store data) are collected from retail stores' checkout scanners from a sample of supermarkets and drugstores. Household panel data (or household data) track purchases of a panel of households on an ongoing basis. Both store and household data are integrated with "causal" information about in-store prices and promotions. Although household data are believed to contain unique information that is not contained in point-of-sales data, they suffer from certain limitations. First, it is believed that households that agree to participate in the panel are a self-

*Anand Bodapati is Assistant Professor of Marketing, Anderson Graduate School of Management, University of California, Los Angeles (e-mail bodapati@ucla.edu). Sachin Gupta is Associate Professor of Marketing, Johnson Graduate School of Management, Cornell University (e-mail: sg248@cornell.edu). The second author acknowledges financial support from the Clifford Whitcomb faculty fellowship. Both authors contributed equally and are listed in alphabetic order. They acknowledge helpful comments from the three anonymous *JMR* reviewers; Jean-Pierre Dubé; and seminar participants at the Wharton School, University of California at Berkeley, Cornell University, University of Chicago, University of Colorado at Boulder, University of Toronto, Northwestern University, and the Marketing Science Conference.

Journal of Marketing Research
Vol. XLI (August 2004), 351–364

selected subset of the random sample of solicited households. The households' noncompliance in recording complete information is also suspected. Consequently, household data are not representative of the purchasing of the population of all households (Bucklin and Gupta 1999). Gupta and colleagues (1996) present evidence of the nonrepresentativeness of panelists. Second, sample sizes are small at the level of an individual market, primarily because of the high cost of setting up and managing household panels. As a result, sampling errors in measurements are large, especially for low-penetration and low-market-share items. Because new products usually fall in this category, low-share items are often of focal interest to management.

Store data typically do not suffer from the problems of nonrepresentativeness and noncoverage to which household data are subject.[1] Such data are readily available to retailers for their own stores and are sometimes shared with manufacturers. For these reasons, academic researchers have been interested in finding applications of store data beyond traditional domains (see, e.g., Bucklin, Russell, and Srinivasan 1998). In this article, we focus on the problem of estimating segmentation structure of consumer response in the form of a latent-class model using store-level aggregate data. Although this model has been extensively applied to household panel data, relatively little is known about its promises and pitfalls in application to store-level data. Our objective is to address this gap in the literature.

RELATIONSHIP TO THE LITERATURE

The latent-class (nested) logit model of brand choice offers an attractive approach to estimate a discrete distribution of consumer preference and responsiveness parameters. Kamakura and Russell (1989) introduced the model, which has become an important analytical tool for the estimation of segmentation structure. A rich stream of articles in marketing has applied this model to household data (e.g., Bucklin and Gupta 1992; Chintagunta, Jain, and Vilcassim 1991; Gupta and Chintagunta 1994).

Motivated by the advantages of store data discussed previously, several articles apply the latent-class model to store data instead of household data (e.g., Besanko, Dubé, and Gupta 2003; Draganska and Jain 2002; Seetharaman 2001; Zenor and Srivastava 1993).[2] Zenor and Srivastava (1993) estimate the model on *aggregated* panel data, which is equivalent in structure to store-level data (we discuss this further in a subsequent section). They validate the model by examining the degree to which implied segment behaviors from the model are congruent with individual household behaviors in the panel data. Seetharaman (2001) proposes a likelihood-based approach to estimate the heterogeneity distributions for both discrete and continuous cases. He uses numerical simulations to demonstrate the feasibility of the approach. Besanko, Dubé, and Gupta (2003) and Draganska and Jain (2002) analyze firms' competitive strategies using the latent-class logit demand model, which they estimate on aggregate data.

Among marketing scientists, there is a wide spectrum of beliefs on the recoverability of segmentation structure from aggregate data. The beliefs range from incredulity, based on the idea that all information on differences among panelists is destroyed upon aggregation, to unquestioning confidence, in which segmentation estimates based on aggregate data are taken at face value and used to construct managerial implications. We find that the correct position is somewhere in between these two extreme beliefs: Recoverability is possible in theory but is quite difficult *in practice*, for reasons we explore herein. To our knowledge, this article provides the first detailed examination of the estimation accuracy of the latent-class model applied to aggregate store-level data. Specifically, our article advances the literature in the area by making the following contributions: First, we establish theoretically that under specific conditions, the parameters of the latent-class nested logit model of brand choice are identified with store data.[3] We provide data requirements for model identification. Thus, our article offers the theoretical basis for models estimated by the researchers listed previously, who *assumed* that the model is identified with aggregate data.

Second, because they are identified, the store data–based maximum likelihood estimates for the latent-class model are consistent. This means that with infinite-sized samples, estimates will have zero mean absolute deviation (MAD). However, the relevant question for managers is, How quickly does the MAD of store data–based estimates diminish with sample size? We provide an answer to this question based on numerical simulations.

Third, our results show that the MAD in the latent-class estimates diminishes much more slowly with store data than with household data. Moreover, the rate is so slow that obtaining estimates with reasonably small MADs often requires unreasonably large sample sizes in store data. We also discuss additional concerns with identification with aggregate data when the household-level model is misspecified or when the model is flexible in terms of response to marketing-mix variables.

The rest of this article is organized as follows: In the next section, we introduce the latent-class nested logit model and review the characteristics of household panel data and store data. We then present theoretical results on identification of the model and subsequently describe a numerical simulation experiment designed to assess the viability of store data and the determinants of performance of store data estimates. We then provide results of the simulation experiments and offer our conclusions.

MODEL DEVELOPMENT

The Nature of Store-Level Data

Consumer purchasing behavior in packaged goods categories has been conceptualized as consisting of three decisions. On each purchase occasion (usually a shopping trip to a given store), the consumer decides (1) whether to buy in the category (purchase incidence), (2) which brand to buy (brand choice), and (3) how much to buy (purchase

[1]In July 2001, Wal-Mart stopped providing point-of-sale data to ACNielsen and Information Resources Inc. This change adversely affected the representativeness of store data and prompted both providers of syndicated data to attempt to improve the coverage of their household panels.

[2]Recent articles in the empirical industrial organization literature also follow this approach. For example, Berry, Carnal, and Spiller (1997) use aggregate data to formulate a two-segment latent-class model to estimate preferences of airline travelers.

[3]For computational ease, we focus herein on the nested logit model. However, we expect that our results translate to the probit model as well.

quantity). Several studies (Bucklin, Gupta, and Siddarth 1998; Chiang 1991; Chintagunta 1993; Gupta 1988) have proposed approaches to model these decisions. The data required to estimate these models are the marketing variables of each alternative in the category on each shopping trip of household panelists and the outcome of the purchasing decision. Furthermore, when household heterogeneity is modeled with panel data, it is necessary to observe a variable that identifies a household uniquely, so that the interdependence between a household's decisions on multiple shopping trips can be modeled. When households are assumed to be homogeneous, the household identifier is not needed.

Consider the case when only store-level data are observed. It is necessary to distinguish between categories in which consumers always (or almost always) purchase at most one unit of the chosen brand (e.g., laundry detergent, ketchup, dish detergent, cookies) and categories in which multiple-unit buying is common (e.g., canned soup, yogurt, canned tuna, carbonated beverages). In this article, we focus on only the former set of categories, namely, ones for which the single-unit assumption is reasonable (Chintagunta [2001] makes the same assumption). Therefore, we limit our attention to the modeling of purchase incidence and brand choice. Without restrictive assumptions, it is not possible to model purchase quantity decisions with only aggregate sales data.

The aggregate data are the *counts* of shopping trips made by all households in a store-week, by purchasing outcome. Because every shopping trip results in a purchase of, at most, one unit of one of J products, the count of purchase trips for each of the J brands, $(N_1, N_2, ..., N_J)$, is the vector of weekly store sales. The count of no-purchase trips is N_0, which is the difference between the total number of trips to the store that week, referred to as "store traffic," and the category sales, $\Sigma_{j=1}^{J} N_j$. In this article, in each store-week, the observed aggregate data are the vector $(N_0, N_1, N_2, ..., N_J)$ and the values of marketing variables.

Note that under the previously described assumption about purchase quantities, because the marketing variable values are the same across all shopping trips in a store-week, the aggregate data as defined previously are a sufficient statistic for the individual-level purchasing outcomes.[4] However, the household identifier variable that links a household's transactions over multiple trips is lost in the aggregation. In other words, under our assumptions, aggregate data are simply household panel data stripped of household identifiers.

Latent-Class Nested Logit Model

We model household incidence and brand-choice decisions using the well-known nested logit model. We present a brief introduction to the model here (for a detailed exposition, see Bucklin, Gupta, and Siddarth 1998; Chintagunta 1992).

In the latent-class model, households are assumed to belong to latent segments indexed by s = 1, 2, ..., S, each of which is characterized by its vector of parameters, Θ_s. As we describe subsequently, the parameters in Θ_s capture the mean frequency of category purchasing, intrinsic preferences for brands, and sensitivity of purchase incidence and brand choices to marketing variables of households in segment s. We define π_s as the probability that a randomly chosen household belongs to segment s, where $0 \le \pi_s \le 1$, and $\Sigma_s \pi_s = 1$.

Conditional on belonging to segment s, $P(inc|X_t, \Theta_s)$ is the probability of purchase incidence in the category on a store trip in week t, given the vector of marketing variables $X_t = (X_{1t}, X_{2t}, ..., X_{Jt})$. Similarly, we define $P(j|inc, X_t, \Theta_s)$ as the probability of choice of brand j, conditional on purchase incidence and membership in segment s, on a store visit in week t. In the nested logit model, the probabilities take the following forms:

$$(1) \qquad P(j|inc, X_t, \Theta_s) = \frac{\exp(\alpha_{js} + \beta_s X_{jt})}{\sum_{k=1}^{J} \exp(\alpha_{ks} + \beta_s X_{kt})},$$

where $\{\alpha_{ks}\}_{k=1}^{J}$ are the intrinsic preferences associated with the J brands for households in segment s, and β_s is the parameter vector associated with the vector of marketing variables X_t. For identifiability, we set the intrinsic preference parameter for the last brand (k = J) in each segment to zero (i.e., $\alpha_{js} = 0 \ \forall$ s).

The incidence probability takes the form

$$(2) \qquad P(inc|X_t, \Theta_s) = \frac{\exp(\gamma_s + \delta_s CV_{st})}{1 + \exp(\gamma_s + \delta_s CV_{st})},$$

where CV_{st} is the category value for segment s in week t, defined as $CV_{st} = \ln[\Sigma_{k=1}^{J} \exp(\alpha_{ks} + \beta_s X_{kt})]$, and is interpretable as the expected utility that a household in segment s derives from purchasing in the category on a trip in week t, under the assumption that the household chooses the brand that maximizes utility (Ben-Akiva and Lerman 1985). It follows that the unconditional choice probability of brand j on a trip in week t is

$$(3) \qquad P(j|X_t, \Theta_s) = P(inc|X_t, \Theta_s) \times P(j|inc, X_t, \Theta_s).$$

Let the response Y = j denote purchase of brand j if j > 0, and let Y = 0 denote nonpurchase in the product category. Furthermore, let p_{jst} be the probability of a consumer from segment s having response Y = j in week t. In terms of the notation already introduced, from Equations 2 and 3, we have the following:

$$(4) \qquad p_{jst} = \begin{cases} 1 - P(inc|X_t, \Theta_s) & \text{for } j = 0 \\ P(inc|X_t, \Theta_s) \times P(j|inc, X_t, \Theta_s) & \text{for } j \ne 0 \end{cases}.$$

Estimation with household panel data. In the household panel data, we observe household h make T_h trips to the store, each of which results in either purchase of a brand in the category of interest or no purchase. The probability of this string of T_h choices occurring, given the household's membership in segment s, is

[4]We exclude consideration of household-specific marketing variables, such as the value of coupons used, from the model. This simplification is necessary because the store model cannot accommodate such variables in a fashion equivalent with the panel data model. It is important to note that such variables are not commonly included even when panel data are used because of concerns with possible endogeneity, in the sense that exposure is not random. Furthermore, in the case of coupons, the researcher typically has no information on coupons available to the household that are *not* redeemed, which causes the specification of the coupon effect to be incorrect.

$$(5)\qquad P(D_h|s) = \prod_{t=1}^{T_h}\prod_{j=0}^{J} p_{jst}^{\delta_{jht}},$$

where D_h represents the string of choices of household h, and δ_{jht} is a 0–1 indicator variable that takes the value of 1 if household h chooses brand j on trip t and the value of 0 otherwise, and j = 0 represents the no-purchase choice. Note that we need the household's identity to construct the probability of the string of choices of household h. The likelihood function for household h, $L(D_h)$, is

$$(6)\qquad L(D_h) = \sum_{s=1}^{S} \pi_s P(D_h|s),$$

and the sample likelihood function is

$$(7)\qquad L(H) = \prod_{h} L(D_h).$$

Estimation with store data. We now proceed to write the likelihood function for the observed store-level data in week t. Because we do not have household identifiers in the aggregate data, we must treat each store visit as independent. Each store visit comes from segment s with probability π_s. Conditional on being from segment s, the likelihood of a particular trip producing response Y = j is p_{jst}, as we noted previously. Therefore, unconditional on segment, the overall likelihood of a particular trip producing response Y = j is $\Sigma_s \pi_s p_{jst}$. Under the assumption that the sample of households that makes a trip to the store is small relative to the population of shoppers, the sampling can be treated as being "with replacement," in the precise sense that Thompson (1992) describes. Therefore, the outcome of each trip is independent of the outcomes of other trips (conditional on model parameters), and the aggregate data are multinomial counts. This implies that the likelihood function for the observed aggregate data $D_{Store,t} = (N_{0t}, N_{1t}, N_{2t}, ..., N_{Jt})$ in week t is

$$(8)\qquad L(D_{Store,t}) = \prod_{j=0}^{J}\left[\sum_{s}\pi_s p_{jst}\right]^{N_{jst}},$$

and the likelihood function for the sample of T weeks is

$$(9)\qquad L(D_{Store}) = \prod_{t} L(D_{Store,t}).$$

An Appendix, available on request, provides a rigorous derivation of Equations 8 and 9.

For both store and household panel data, we estimate the unknown parameters using the maximum likelihood method. We determine the number of segments S by carrying out the estimation for models with an increasing number of segments until there is no significant improvement in model fit between the S segments and the S + 1 segment solutions. The model that is often picked as "best" is the one that minimizes a score that trades off model complexity for model fit (e.g., the Bayesian information criterion [BIC]; see Schwarz 1978).

IDENTIFICATION OF THE MODEL WITH STORE DATA: POSSIBILITY AND PRECISION

We begin by demonstrating that the nested logit model of purchase incidence and brand choice is identified with both store data and household data but that the data requirements are more severe for store data. We then discuss concerns with identification in empirical settings. Subsequently, we discuss the implications of finite sample size for estimation error in parameter estimates from store data.

Requirements for Identifiability

As we noted previously, we restrict attention to the modeling of purchase incidence and brand choice decisions alone. With regard to purchase quantities, we assume that each purchase consists of one unit of the product only. Furthermore, we assume that a latent-class nested logit model of purchase incidence and brand choice with S segments, as we described previously, is correctly specified for the data. We discuss the potential effects of model misspecification in a subsequent section.

The results on identifiability may be formally stated as the three lemmas we provide subsequently. Lemma 1 is a classic result that can be found in graduate-level statistics textbooks, such as that of Gourieroux and Monfort (1995). Lemmas 2 and 3 are results we developed.

We denote the likelihood function for some stochastic data y in terms of parameters θ and conditional on some predictors x by $L(y; \theta, x)$. We denote the expected Fisher information matrix that corresponds to this likelihood function by $I_y(\theta, x)$:

$$(10)\qquad I_y(\theta_0, x) = -E_y \nabla_\theta \nabla_\theta^T \log L(y; \theta, x)\Big|_{\theta=\theta_0},$$

where θ_0 is the true value of the parameter θ, and ∇_θ is the gradient of the likelihood function with respect to θ.

In the context of the problem addressed by this article, θ represents the parameters of the nested logit latent-class model, and x represents the store's marketing-mix levels for the various brands in the weeks spanned by the store data. Let there be M marketing-mix variables. Assume that the number of weeks spanned by the data is T, and the vector of marketing-mix levels in each week is denoted by x_t, so that $x = \{x_t\}_{t=1}^{T}$. There are J brands with M marketing-mix variables measured for each brand, so each x is a vector in $\Re^{JM}$ space. We consider estimating the latent-class model on the basis of either a household panel data set or a store data set. In the case of a household data set, the likelihood is as defined in Equation 7. In the case of a store data set, the corresponding likelihood is as defined in Equation 9.

We use the notation $A \succ 0$ to denote that matrix A is positive definite. We denote the expected Fisher information matrix by $I_S(\theta_0, x)$ for the store data and by $I_H(\theta_0, x)$ for the household data. We are now in a position to state our three lemmas.

Lemma 1: If $I_y(\theta_0, x) \succ 0$, the parameter θ is locally identified by the likelihood function $L(y; \theta, x)$.

Lemma 2: Assume that the store data span T weeks and that the marketing-mix vectors $x_1, x_2, ..., x_T$ are independent draws from densities with support on the entire vol-

ume of $\Re^{JM}$. If $T > \{[S(J + M + 2)S - 1]/J\}$, then $I_s(\theta_0, x) \succ 0$, almost surely.

Lemma 3: Assume that the household panel data span T weeks and that the marketing-mix vectors $x_1, x_2, ..., x_T$ are independent draws from densities with support on the entire volume of $\Re^{JM}$. If $T > \{\log[S(J + M + 2)]/\log(J + 1)\}$, then $I_H(\theta_0, x) \succ 0$, almost surely.

(For the proof of Lemma 1, see Gourieroux and Monfort 1995. The proofs of Lemmas 2 and 3 are directly based on the ranks of the information matrices obtained with T weeks of data in the two cases; details are in an Appendix available on request from the authors.) Lemma 1 asserts that if the expected information matrix is positive definite, the parameters are locally identified. In turn, Lemmas 2 and 3 identify the conditions under which the information matrices in the store and household data are positive definite. Collectively, the three lemmas give the conditions (in terms of the minimum number of weeks of data required) under which the latent-class model is locally identified. Lemmas 2 and 3 also show that the threshold number of weeks required for identification grows linearly with the number of segments for store data but grows only logarithmically for household data. In this sense, the data requirement for estimation of the latent-class model is considerably more severe for store data than for household data.

Concerns with Identification in Practice

In the foregoing discussion of identification with aggregate data, we assume that the household-level model is correctly specified (i.e., it is a homogeneous nested logit model). If the model is misspecified, model identification is not a meaningful idea in theory. However, this case is of practical interest. Suppose that the households are homogeneous. We expect that if the household model is misspecified, store data will show evidence of heterogeneity when none exists. For example, if the true household choice model is probit and all households have the same probit parameter vector so that there is no across-household heterogeneity, a latent-class logit model on aggregate data will incorrectly infer multiple segments and conclude that there *is* across-household heterogeneity. This is because a mixture of logits can closely approximate a probit model (McFadden and Train 2000). Similarly, suppose that the choices of each household are generated by a mixture of logits because of intrahousehold heterogeneity (Seetharaman, Ainslie, and Chintagunta 1999), though households are all alike. If the household-level model is misspecified as homogeneous logit, a latent-class model based on aggregate data may incorrectly suggest that there are segments of households in the market. The tendency of the latent-class model to find across-household heterogeneity when none exists (as in these two situations) is not specific to the logit model or to its independence-from-irrelevant-alternatives property; rather, it comes from the misspecification. Indeed, the same tendency would arise if we used some other model for the household-level choice that was also misspecified. Take the first situation mentioned previously but with the roles of probit and logit reversed, so that all households' choices came from the same logit model, and a latent-class probit model is estimated on aggregate data: We would incorrectly infer multiple segments and conclude that there is across-household heterogeneity. This happens because a mixture of probits can closely approximate a logit model. The probit model does not have the independence-from-irrelevant-alternatives property; however, here the latent-class model estimated with aggregate data also tends to overstate the extent of heterogeneity.

Another practical concern with identification with aggregate data arises when the true household-level model is flexible in terms of response to marketing-mix activities. In this situation, we expect that even if the model estimated with aggregate data is correctly specified, it will be difficult to identify the heterogeneity distribution. Given the hypothesized choice model at the household level, homogeneity across households implies a certain pattern of market share changes in response to marketing-mix changes. If the observed changes in market shares are discrepant from the changes implied by homogeneity, the latent-class model ascribes them to heterogeneity and uses the discrepancies to estimate the parameters of the latent-class model. This is a key point: The stronger the discrepancies, the better is the recoverability of segmentation structure from aggregate data. Similarly, the weaker the discrepancies, the poorer is the recoverability. With a flexible household-level model specification, the discrepancies between observed market share movements and ones implied by homogeneity are likely to be weak, because the homogeneous model may be adaptive enough to fit even complex aggregate share movements well. This would make it difficult to identify the heterogeneity distribution from aggregate data. In the extreme situation of the household-level model being a mixture of logits, the homogeneous model would be so flexible and its discrepancies with observed share movements so small that we anticipate that the latent-class distribution across households will be poorly recovered with aggregate data.

Estimation Error in Finite Samples

The local identifiability of the latent-class model from store data is good news for the recoverability of segmentation structure from aggregate data, because identifiability also ensures consistency. This means that with infinite-sized samples, we will obtain estimates that have zero error, where "error" can be measured in several ways. However, we consider error the MAD. As a practical matter, the sample size in real data is never infinite. Accordingly, we now consider asymptotic theory that is informative of the *finite sample* behavior of the maximum likelihood estimate $\hat{\theta}$ for the parameter θ, which takes the value of θ_0 for the data generation process. We consider each parameter in the vector $\hat{\theta}$ elementwise. We denote the ith component of θ as θ_i. In the subsequent passage, we suppress the dependence of the estimators on the predictors x. The MAD for θ_i is $E|\hat{\theta}_i - \theta_{0i}|$, where the expectation is taken over the sampling distribution for the estimate $\hat{\theta}_i$. Let n denote the sample size, which we take here to be the number of panelists. A consequence of the theory discussed by Pace and Salvan (1997, Chap. 6) is that an expression for the MAD of maximum likelihood estimators can be written as follows:

$$\text{MAD}(\theta_{0i}) = \frac{D_1(\theta_{0i})}{\sqrt{n}} + \frac{D_2(\theta_{0i})}{n} + O\left(\frac{1}{n^{3/2}}\right). \tag{11}$$

The term $0(1/n^{3/2})$ abstracts all terms in which n appears with exponent of –3/2 or less. In the text that follows, we drop the subscript i, so equations of the previous form should be interpreted as applying elementwise to the components of the parameter vector θ. The factors $D_1(\theta)$ and $D_2(\theta)$ do not depend on the sample size but do depend on the choice model and the exogenous predictor variables. Closed-form expressions for $D_1(\theta)$ and $D_2(\theta)$ can be formulated in terms of the cumulants of the derivatives of the log-likelihood (Shenton and Bowman 1977). It turns out that $D_1(\theta)$ is dominated by the inverse expected Fisher information, the consequence of which is that $D_1(\theta_{0i})/\sqrt{n}$ is usually the asymptotic standard error for the estimate. Computation of the terms $D_1(\theta)$ and $D_2(\theta)$ directly from the closed-form expressions is difficult, particularly when there are exogenous predictor variables involved. The more common approach is to compute $D_1(\theta)$ and $D_2(\theta)$ indirectly (as we do herein) with the empirical values of MAD obtained from a simulation design.

As Equation 11 makes clear, the MAD decreases with the sample size. For infinite sample sizes, the value of MAD goes to zero, as is to be expected for consistent estimators. However, to the manager, of more practical interest is the question, How quickly does the MAD diminish with sample size? The answer comes from noting that the convergence rate depends to first order on $D_1(\theta)$ and to second order on $D_2(\theta)$. Because $\sqrt{n}$ increases at a considerably slower rate than n, the first-order asymptotic effects dominate the second-order effects for even moderate values of sample size. For this reason, the rest of our discussion focuses on $D_1(\theta)$ rather than $D_2(\theta)$. In the next section, we compare the accuracy of estimates from panel data and from store data by comparing their corresponding value of $D_1(\theta)$ for the various cells in a simulation design.

NUMERICAL EXPERIMENTS

In this section, using numerical simulations, we assess the magnitude of the error in estimates of the latent-class nested logit model parameters on the basis of household data and store data. Our simulation design strategy is similar to that of Andrews, Ainslie, and Currim (2002), who examine the validity of different methodological approaches to capture heterogeneity in choice data. We identify six factors (see Table 1) based on theory and a priori experimentation that we expect will affect the error in estimates. The choice of levels of the factors is guided by our desire to have a realistic range of variation. In each cell of the simulation design, we generate data for multiple sample sizes and sample replicates.

As we discussed in the previous section, the theory in Lemmas 1–3 ensures that the store data–based estimates converge to the true parameter values with infinite samples, but of greater managerial interest is the finite sample behavior of the estimates and the rate at which the estimates converge. By studying how the estimation error diminishes with sample size, we can draw inferences from the convergence rate. We generate data for 22 different values of sample size n, which indicates the number of panelists. The values are located at equal intervals in log space: 350, 420, 500, 590, 710, 840, 1000, 1190, 1410, 1680, 2000, 2380, 2830, 3360, 4000, 4760, 5660, 6730, 8000, 9510, 11,310, and 13,450. For each of the 22 values of n, we generate 28 sample replicates to compute the empirical MAD of the estimates. We regress the empirical MAD on $1/\sqrt{n}$ and 1/n to obtain estimates of $D_1(\theta)$ and $D_2(\theta)$. Knowing these two factors enables us to evaluate the convergence rate of the latent-class estimates for the store data set and to compare it with that for the household data set.

In each cell of the experiment, we estimate 1232 models (22 values of n × 28 sample replicates × two types of data [panel and store]). Because this estimation is computationally expensive, we develop a parsimonious experiment design. As is described in Table 1, we focus on two and three segments in the heterogeneity distribution and on two and three brands in the choice set (Factors 5 and 6). We manipulate price variation (Factor 1) at low and high levels by varying the frequency and depth of price discounts. For example, at the low level, each brand is at regular price 70%

Table 1
SIMULATION DESIGN FACTORS

Factor	*Level 1*	*Level 2*	*Level 3*	*Remarks*
1. Price variation (PRICEVAR)	Low	High		Two parameters are varied to manipulate price variation: frequency of price discounts and depth of price discounts.
2. Presence of binary marketing variable (DISPLAY)	No	Yes. Coefficient is set at 1 for each segment.		This variable represents feature advertising or in-store display. For each brand, it is correlated negatively with price.
3. Separation between segments (HETERO)	Low (1.5, –1), (–1, 1), (–1, –1)	Medium (2, –1), (–1, 1), (–1.5, –1)	High (2.5, –1), (–1, 1), (–2, –1)	The distance between the brand-specific constants for Brands A and C is increased in Segments 1 and 3, respectively. Note that the brand-specific constant for Brand C is zero. $(\alpha_{A1}, \alpha_{B1})$, $(\alpha_{A2}, \alpha_{B2})$, $(\alpha_{A3}, \alpha_{B3})$
4. Dependence of purchase incidence on prices (DELTA)	Low: $\delta_1 = \delta_2 = \delta_3 = .3$	High: $\delta_1 = \delta_2 = \delta_3 = .5$		
5. Number of brands (NUMBRND)	Two	Three		In the two-brand case, Brand B is considered absent.
6. Number of segments (NUMSEG)	Two: $(\pi_1, \pi_2) = (.3, .7)$	Three: $(\pi_1, \pi_2, \pi_3) = (.2, .3, .5)$		In the two-segment case, Segment 2 is considered absent.

Notes: We did not manipulate the following parameters in the experiment: segmentwise price parameters $(\beta_1, \beta_2, \beta_3) = (–1, –2, –3)$ and segmentwise intercepts in the utility of category purchasing $(\gamma_1, \gamma_2, \gamma_3) = (–1.5, –.5, .5)$.

of the time; when discounted, each brand offers six levels of price discounts ranging from 10% to 35%. Prices of competing brands are negatively correlated, which reflects that the retailer manages the category to promote one brand at a time. At the high level of price variation, prices are discounted more deeply and more often. Factor 2 is the presence of a binary marketing variable that represents in-store displays or feature advertising. This activity tends to coincide with price discounting, and the variable is accordingly correlated with prices. An important factor identified in previous research (e.g., Andrews, Ainslie, and Currim 2002) is the extent of heterogeneity, or separation, between segments (Factor 3). In our design, we accomplish this by manipulating the distance between the brand-specific constants of different segments. The final factor (Factor 4) is the degree of dependence between the purchase incidence and the brand-choice decisions. We manipulate this through the coefficient of the inclusive value in the nested logit model.

To generate the household data, we simulate 100 vectors of marketing-mix levels to correspond with 100 store-weeks. Each panelist household makes a store visit on a random subset of 50% of the weeks. Consistent with the latent-class model, a panelist belongs to one of S segments, and the panelist's vector of preference and responsiveness parameters is fixed according to segment membership. On each store visit, we generate a panelist's purchasing outcome under the following assumptions: (1) Given the values of marketing-mix variables, we generate a discrete choice from a nested logit model. A realized choice is either no purchase or purchase of a brand in the category. (2) The purchase quantity is at most one unit. These discrete choices constitute the household panel data. The discrete choices of all households in each of the 100 store-weeks are counted to yield aggregated panel or store data.

Note that the estimated models are always correctly specified in terms of the functional form of the model, explanatory variables, and number of segments, both for the household data and for the store data. As a result, we can draw inferences about the effects of aggregating the household data to store data, without the confounding effects of model misspecification. Subsequently, we speculate on the possible effects of model misspecification, but we leave a rigorous investigation of this question to further research.

In this design, the average number of shopping trips per customer is 50, which implies that on the high end of our range of sample sizes (i.e., 13,450), the total number of transactions in any one replication of the choice data set is 672,500, which far exceeds the annual number of shopping visits to a typical grocery supermarket.[5]

RESULTS AND INTERPRETATIONS

We now discuss results from the numerical simulations. We focus on recovery of segment-level parameter estimates from store data compared with that of household data. We examine overall measures of estimation accuracy and variation in estimation accuracy as a function of the six factors in the simulation experiment. Subsequently, we assess estimation accuracy of the mean (across-segment) demand parameters.

[5]We refer readers to two publications of the Food Marketing Institute (2002a, b) for statistics on store traffic and transaction volume.

Overall Estimation Accuracy of Segment-Level Parameters

Pooled across all the cells in the simulation design, there are 21 parameters in the models (for a description and listing of parameters, see Tables 1 and 2). To facilitate comparisons across parameters, and across cells when the parameter values differ, we define measures of estimation accuracy relative to true parameter values. Specifically, we define relative MAD as $E|(\hat{\theta}_i - \theta_{0i})/\theta_{0i}|$. The relative MAD is interpretable as the percentage deviation from the true parameter value. We estimate the MAD by (1) computing the relative absolute deviations for each of the 28 replicates in any given simulation cell for a given sample size, (2) deleting the highest three values and the lowest three values to reduce the influence of outliers, and (3) computing the average of the remaining 22 values.

In Figure 1, we show the relative MAD of store data estimates and household data estimates for each of the 22 sample sizes. The numbers displayed are simple averages across all parameters and across all cells in the simulation design. Figure 1 reveals that the error in store data estimates is substantially larger than it is in household data estimates, even at the high end of sample size. Across the 22 sample sizes studied, the relative MAD is, on average, 11 times larger. Thus, there is substantial loss in accuracy due to the use of store data instead of household data. Furthermore, as we expected, the accuracy of estimates from both store and household data improves with sample size. However, improvements accrue more quickly with household data than with store data. As a result, the ratio of error in store estimates to error in household data estimates increases with sample size (this result is not shown).

To provide a more detailed picture of the magnitude of the errors, in Tables 2 and 3, we show descriptive statistics of the relative MAD for each of 21 parameters for the smallest (n = 350) and largest (n = 13,450) sample sizes in our data. In the upper part of Tables 2 and 3, we show the average and standard deviation of the relative MAD across all cells in the simulation design. In the lower part of Tables 2 and 3, we show the average and standard deviation of the relative MAD value across all parameters for a certain factor level in the simulation design. The detailed results reiterate our previous finding of large deterioration in the accuracy of parameter estimates from store data compared with that of panel data for each of the parameters. We note that at the largest sample size of 13,450 panelists, the relative MADs are quite small for household data estimates; the largest mean error is only 5.8%. In contrast, for store data, the errors remain intolerably large: The smallest average error is 5.5%, and the largest is 112%. Notably, in general, the average error in store data estimates at the largest sample size is greater than the average error in household data estimates at the smallest sample size.

Effect of Simulation Factors on Estimation Accuracy

To explore the effects of the six simulation factors on the estimation accuracy of store data estimates, we estimate an analysis of variance model with relative MAD as the dependent variable. For this analysis, we pool the estimates across parameters and across sample sizes. Consequently, in addition to the factor effects, we include fixed effects for each of the 21 parameters and for each of the 22 sample sizes. Although our investigation revealed that several inter-

Table 2
RELATIVE MAD (%) IN HOUSEHOLD AND STORE DATA: SMALL SAMPLE SIZE (N = 350)

	Household Data		*Store Data*	
Parameters	*Mean*	*Standard Deviation*	*Mean*	*Standard Deviation*
π_1	.4	.6	62.0	23.8
π_2	15.7	21.6	57.8	18.1
π_3	4.8	10.2	33.7	12.2
α_{A1}	3.0	1.2	77.3	39.9
α_{A2}	8.2	5.6	102.8	58.4
α_{A3}	4.3	5.2	46.8	41.5
α_{B1}	10.0	4.0	201.0	141.2
α_{B2}	4.2	1.6	94.1	32.9
α_{B3}	3.5	1.5	36.8	34.2
β_1	10.7	4.1	95.2	52.6
β_2	7.4	5.9	43.5	15.9
β_3	3.5	4.5	24.4	22.0
$\beta_{dis,1}$	9.6	3.9	62.6	41.5
$\beta_{dis,2}$	10.4	8.3	59.5	30.0
$\beta_{dis,3}$	5.6	4.6	34.3	43.4
γ_1	6.2	3.4	73.7	42.1
γ_2	35.8	36.7	191.3	88.4
γ_3	22.0	14.1	188.2	107.7
δ_1	19.3	13.2	91.0	67.8
δ_2	13.3	8.4	59.5	38.8
δ_3	6.4	3.8	25.7	21.2

	Household Data		*Store Data*	
Simulation Factor Levels	*Mean*	*Standard Deviation*	*Mean*	*Standard Deviation*
HETERO: low	8.9	11.2	81.9	79.1
HETERO: medium	9.5	15.1	74.5	71.0
HETERO: high	9.3	11.7	73.8	73.0
PRICEVAR: low	10.2	13.1	78.7	78.9
PRICEVAR: high	8.3	12.4	74.8	69.9
DELTA: low	7.4	8.3	76.5	76.3
DELTA: high	11.1	15.9	77.0	72.8
NUMSEG: two	6.4	7.2	58.1	60.4
NUMSEG: three	11.1	15.1	89.2	80.3
NUMBRAND: two	12.0	16.5	83.9	75.9
NUMBRAND: three	6.9	7.7	70.7	72.8
DISPLAY: no	10.1	13.7	85.3	82.5
DISPLAY: yes	8.5	12.0	69.5	66.2

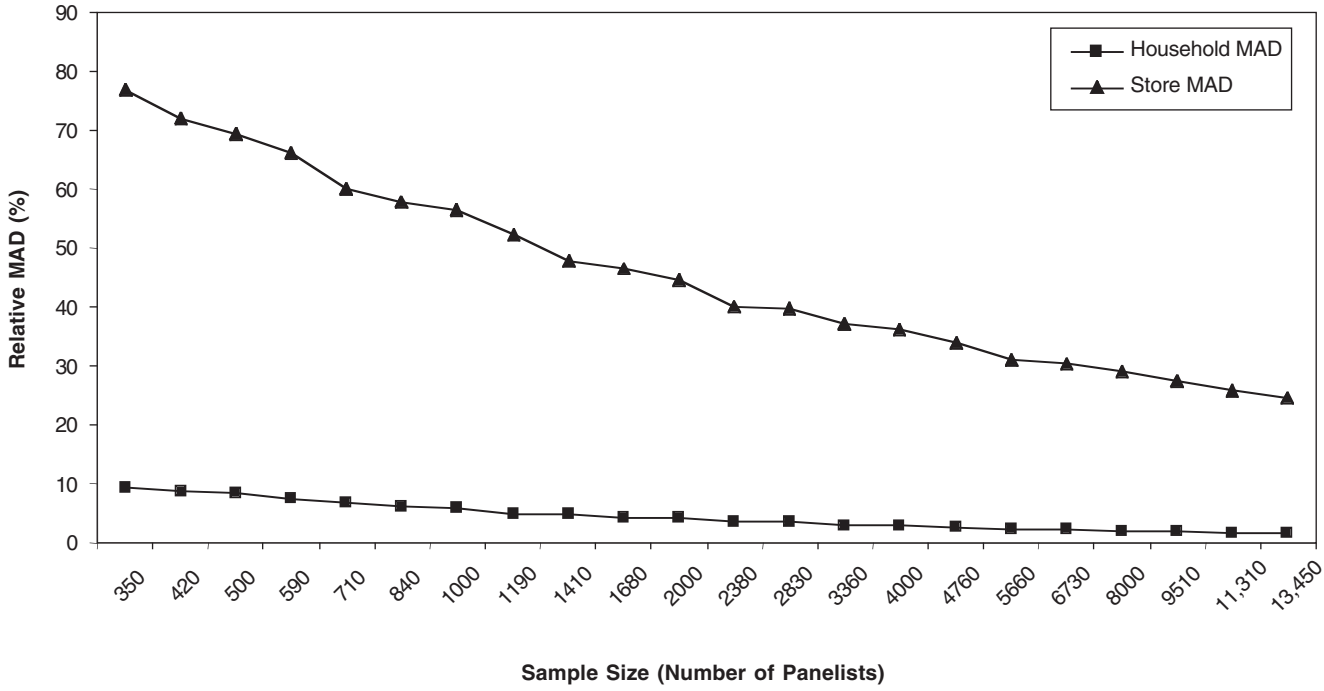
action effects between the six factors were significant, because our analysis of the simulation factors is primarily exploratory, we present (in Table 4) and discuss the results only from the model without interactions. All six main effects are significant ($p < .02$). The relative magnitudes of the F-statistics indicate the relative explanatory power of each of the six factors in the context of our simulation design. We find that the number of segments and the number of brands are the two most important factors; the degree of heterogeneity and the dependence of purchase incidence on prices have low explanatory power.

From the main effects of each of the six factors (we do not show these results for reasons of space), we find that the accuracy of store data estimates is enhanced when there is greater choice heterogeneity, higher price variation, more brands, and a binary display variable. These directions seem to be intuitive because each factor implies that more information becomes available in the store data to identify the segment-level choice models. For example, with three brands rather than two brands, we have three prices in each week instead of two. On the flip side, the accuracy of store data estimates decreases with a greater number of segments. This is because an increased number of segments implies that there are more parameters in the model, thereby increasing the difficulty of the estimation task while holding the data constant (we explain this further in a subsequent section). The effect of the degree of dependence of purchase incidence on prices is negligible but statistically significant.

Accuracy of Mean Versus Segment-Specific Effects

In Figures 2 and 3, using store data, we show the relative MAD in the estimate of the mean (across-segments) price parameter and segment-specific price parameters for each of 22 sample sizes. We show results separately for the case of two- versus three-segment models. The mean price estimate is the average of segment-specific price effects, weighted by segment sizes. Several observations emerge from Figures 2 and 3. First, the relative MAD in the mean estimate is low compared with errors in the segment-specific price effects. In particular, in the three-segment model, the error in the mean price effect is no greater than the error in any of the segment-specific effects. The smaller error in the mean effect may be due to the netting out of errors in segment-specific estimates in the process of averaging. Second, the relative MAD in the mean effect in the three-segment case is greater than it is in the two-segment model, which is consistent with the previous finding that, in

Figure 1
RELATIVE MAD (%) IN STORE AND HOUSEHOLD DATA ESTIMATES



general, errors increase with number of segments. Finally, we note that errors are smaller for larger segments. In the two-segment case, the relative sizes of Segments 1 and 2 are .3 and .7, and the relative MAD is higher for Segment 1 at all sample sizes. We observe the same phenomenon in the three-segment case.

The Behavior of the Convergence Rate $D_1(\theta)$

As we discussed previously, because we vary sample size over a wide range, we can estimate the value of the asymptotic convergence rate factor, $D_1(\theta)$, that we refer to in Equation 11. For the analysis presented here, we take the sample size n to be the number of transactions rather than the number of panelists. Bear in mind that the number of transactions is, on average, 50 times greater than the number of panelists. We also divide the $D_1(\theta)$ by the true parameter value so that the resulting quantity, which we refer to as the relative $D_1(\theta)$, gives the rate at which relative MAD reduces with sample size. In Tables 5 and 6, for each model parameter, we report the value of relative $D_1(\theta)$ averaged over all cells that match a certain level for a certain factor. Tables 5 and 6 report this average value for every level of every factor separately for household and store data. As with Tables 2 and 3, given the way that we constructed each table, the row totals (rightmost column) in each table are the simple average relative $D_1(\theta)$ across all cells for each parameter. Similarly, the column totals in each table are the average relative $D_1(\theta)$ across all parameters for a certain factor level in the simulation design.

We now present some calculations that illustrate use of the asymptotic factors reported in Tables 5 and 6. Consider the number that appears in Table 6 and corresponds to HETERO = high for parameter π_2. This suggests that if we observe n transactions (store visits) in the store data set, and if the heterogeneity factor is set at "high," the relative MAD for parameter π_2 will be $178.08/\sqrt{n}$. Therefore, if the store has 100,000 transactions, on average, the relative MAD will be $178.08/\sqrt{100{,}000} = 56.3\%$, which is sizable. Another possible question is, What is the transaction volume that is needed in store data to achieve a relative MAD of 5%? We obtain the answer by applying the converse interpretation of relative $D_1(\theta)$: $n = (178.08/.05)^2 = 12{,}684{,}995$, which is a transaction volume rarely observed in typical grocery supermarkets in the period of a year. In contrast, the corresponding value of relative $D_1(\theta)$ for household data is 18.28, so the required transaction volume to achieve the same target of 5% is $n = (18.28/.05)^2 = 133{,}663$, which is a smaller number by a factor of almost 100.

Comparison of the household factors (Table 5) with the store factors (Table 6) shows systematically that the store factors are much larger than the household factors. The more significant message from Table 6 is that the convergence rate with store data estimates is very slow indeed. The illustrative calculation in the previous paragraph took a value of relative $D_1(\theta)$ that is typical of Table 6 and demonstrated that unreasonably high transaction volumes are needed to obtain reasonably accurate estimates of the parameters. In addition, Tables 5 and 6 reinforce the simulation factor effect that we already identified, in that the same factors shown to be influential in the analysis of variance of relative MAD also are influential on relative $D_1(\theta)$.

CONCLUSION AND FUTURE RESEARCH AREAS

Researchers in marketing and economics have recently begun using aggregate store-level data to estimate segmentation structures based on latent-class models of brand

Table 3
RELATIVE MAD (%) IN HOUSEHOLD AND STORE DATA: LARGE SAMPLE SIZE (N = 13,450)

	Household Data		*Store Data*	
Parameters	*Mean*	*Standard Deviation*	*Mean*	*Standard Deviation*
π_1	.1	.1	23.2	9.5
π_2	3.1	4.5	21.7	7.6
π_3	.9	2.1	12.9	5.8
α_{A1}	.6	.2	26.7	13.3
α_{A2}	1.5	1.0	34.1	27.2
α_{A3}	.6	.5	12.6	9.6
α_{B1}	1.7	.7	35.1	35.4
α_{B2}	.7	.3	38.9	13.4
α_{B3}	.7	.3	12.6	13.3
β_1	1.8	.8	27.1	21.2
β_2	1.3	1.0	14.5	8.0
β_3	.5	.4	5.5	3.6
$\beta_{dis,\ 1}$	1.6	.8	15.3	9.0
$\beta_{dis,\ 2}$	1.7	.9	20.2	16.4
$\beta_{dis,\ 3}$	1.0	.6	6.1	6.7
γ_1	1.1	.6	24.2	11.1
γ_2	5.8	5.1	112.1	52.4
γ_3	3.4	2.4	55.5	31.8
δ_1	2.9	1.9	22.2	11.4
δ_2	2.0	1.0	24.7	19.9
δ_3	.9	.4	7.0	4.8

	Household Data		*Store Data*	
Parameters	*Mean*	*Standard Deviation*	*Mean*	*Standard Deviation*
HETERO: low	1.5	2.2	25.2	27.0
HETERO: medium	1.5	2.1	24.6	26.4
HETERO: high	1.5	1.8	24.1	28.6
PRICEVAR: low	1.7	2.3	26.7	28.4
PRICEVAR: high	1.3	1.7	22.6	26.1
DELTA: low	1.3	1.4	24.9	27.3
DELTA: high	1.8	2.6	24.4	27.4
NUMSEG: two	1.0	1.1	16.3	15.4
NUMSEG: three	1.9	2.4	30.2	31.8
NUMBRAND: two	1.9	2.7	27.3	26.6
NUMBRAND: three	1.2	1.2	22.4	27.7
DISPLAY: no	1.6	2.3	30.5	32.9
DISPLAY: yes	1.4	1.8	19.7	20.2

Table 4
F-STATISTICS OF MAIN EFFECTS OF FACTORS ON ACCURACY OF STORE DATA ESTIMATES

Effects	*Relative MAD (Store)*
Fixed Effects	
Parameters	1119.99
Sample size	279.33
Main Effects of Factors	
Heterogeneity	8.35
Price variation	167.75
Delta	5.53
Number of segments	1757.46
Number of brands	722.25
Display	423.37

Notes: All tests are significant at $p < .02$; $R^2 = .51$, and N = 31,860.

choice. However, both theoretical and practical questions about identification of the model parameters with store data remain unanswered. The purpose of this article is to investigate these questions.

We show that the parameters of the latent-class model are theoretically identified with store data. Thus, infinite-sized samples would yield consistent parameter estimates with zero MAD. To study finite sample properties of the estimates, we simulate almost 60,000 store data sets that span various market scenarios. On the basis of this simulation study, we find that the bias in parameter estimates from store data is large for realistic sample sizes. Moreover, the reliability of estimates is so low that in a single sample, as is the case in real data, it is likely that the estimated parameters are far from their true values.

Why Is It So Difficult to Recover Segmentation Structure from Aggregate Data?

As we pointed out previously, the only information in aggregate data about heterogeneity lies in discrepancies in aggregate share movements relative to movements in shares the model would predict if there were no heterogeneity. Any effort to recover segmentation succeeds only to the extent that the discrepancies can be ascribed to a heterogeneous latent-class model. If there are no strong discrepancies in the first place, there is not much information to exploit to parameterize the latent-class model. The absence of strong discrepancies can arise in at least two situations. The first situation is when the household-level model is so flexible

Figure 2
RELATIVE MAD (%) IN MEAN AND SEGMENT-SPECIFIC PRICE PARAMETERS IN STORE DATA: TWO-SEGMENT CASE

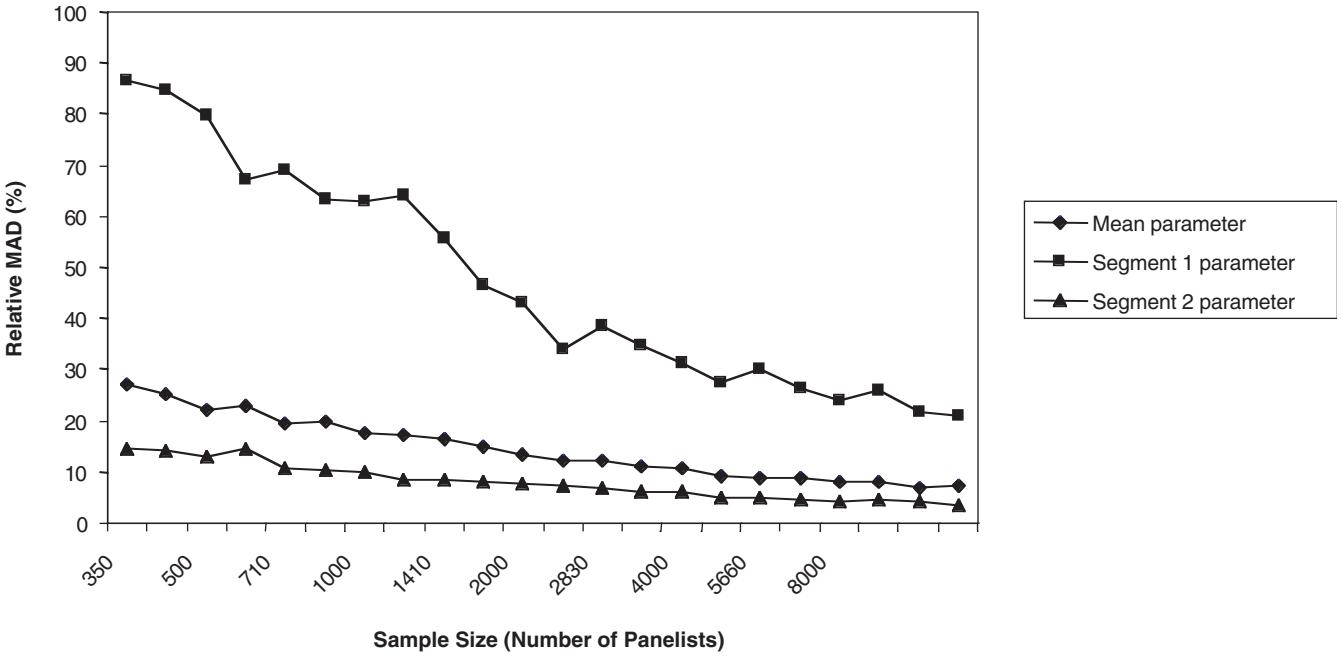
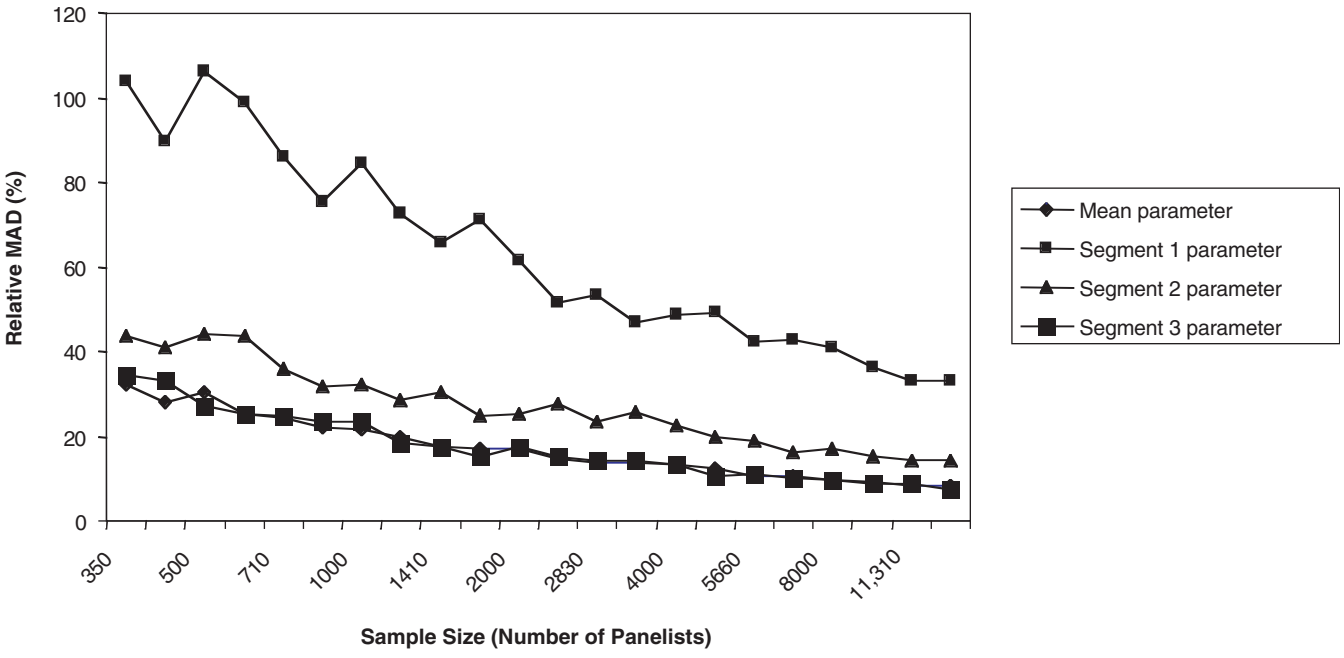


Figure 3
RELATIVE MAD (%) IN MEAN AND SEGMENT-SPECIFIC PRICE PARAMETERS IN STORE DATA: THREE-SEGMENT CASE



that the homogeneous model is adaptive enough to fit even complex aggregate share movements well. The second situation is when the sample size is small. In small sample sizes, there is considerable variation in observed market share as a result of sampling error. Discrepancies in observed market share (compared with the expected shares when there is homogeneity) attributable only to heterogeneity and not sampling error become much more difficult to detect. In turn, this makes it difficult to estimate latent-class models from aggregate data when the sample sizes are

Table 5
RELATIVE D FACTOR: HOUSEHOLD DATA ESTIMATES

	HETERO			*PRICEVAR*		*DELTA*		*NUMSEG*		*NUMBRAND*		*DISPLAY*		
Parameters	*Low*	*Medium*	*High*	*Low*	*High*	*Low*	*High*	*Two*	*Three*	*Two*	*Three*	*No*	*Yes*	*Mean*
π_1	1.21	.38	.09	.45	.67	.44	.68	.39	.73	.69	.43	.56	.56	.56
π_2	26.87	31.77	18.28	29.31	21.97	6.95	44.33		25.64	48.43	2.85	27.43	23.85	25.64
π_3	8.27	9.59	5.49	8.87	6.7	2.14	13.43	.17	15.4	14.62	.95	8.32	7.25	7.78
α_{A1}	4.94	3.94	3.49	4.16	4.09	4.23	4.02	3.68	4.57	4.08	4.17	3.93	4.31	4.12
α_{A2}	10.4	12.54	12.52	11.48	12.16	7.91	15.72		11.82	14.18	9.45	12.52	11.12	11.82
α_{A3}	5.71	5.1	4.7	5.03	5.31	3.52	6.81	3.75	6.59	6.17	4.17	5.19	5.15	5.17
α_{B1}	10.93	13.74	15.43	13.44	13.29	13.48	13.25	11.74	14.99		13.37	13.56	13.17	13.37
α_{B2}	5.63	5.97	6.14	6.2	5.63	5.33	6.49		5.91		5.91	5.73	6.1	5.91
α_{B3}	4.99	5.31	5.41	5.54	4.94	4.07	6.4	4.41	6.07		5.24	5.12	5.35	5.24
β_1	13.85	14.38	15.47	17.19	11.94	15.22	13.91	13.25	15.88	14.6	14.53	14.07	15.06	14.57
β_2	11.73	10.97	8.67	12.1	8.82	6.36	14.56		10.46	15.26	5.66	10.57	10.34	10.46
β_3	3.94	4.45	4.7	4.68	4.05	2.92	5.81	2.96	5.76	5.92	2.8	4.36	4.37	4.36
$\beta_{dis,1}$	11.97	12.08	13.75	13.89	11.31	14.24	10.96	11	14.2	12.88	12.32		12.6	12.6
$\beta_{dis,2}$	12.69	15.14	11.09	14.06	11.88	10.73	15.21		12.97	16.53	9.41		12.97	12.97
$\beta_{dis,3}$	8.07	7.9	7.85	9.3	6.58	6.87	9.01	6.17	9.72	9.72	6.16		7.94	7.94
γ_1	7.14	8.05	10.7	10	7.27	7.23	10.04	7.87	9.4	8.41	8.85	9.11	8.15	8.63
γ_2	55.44	53.53	41.16	56.67	43.42	28.2	71.88		50.04	78.02	22.06	56.18	43.91	50.04
γ_3	28.46	31.03	26.78	34.2	23.31	21.63	35.89	21.25	36.27	30.56	26.95	31.16	26.36	28.76
δ_1	26.8	24.21	23.92	29.64	20.32	30.57	19.39	22.21	27.74	24.89	25.06	33.76	16.2	24.98
δ_2	18.07	16.42	18.14	18.56	16.53	17.49	17.6		17.54	21.29	13.8	20.8	14.29	17.54
δ_3	7.67	7.99	7.93	9.07	6.66	7.91	7.82	6.52	9.21	7.78	7.95	8.59	7.15	7.87
Mean	12.79	13.1	11.88	14.23	10.95	10.17	15	8.22	15.5	16.13	9.6	14.08	11.33	

Table 6
RELATIVE D FACTOR: STORE DATA ESTIMATES

	HETERO			*PRICEVAR*		*DELTA*		*NUMSEG*		*NUMBRAND*		*DISPLAY*		
Parameters	*Low*	*Medium*	*High*	*Low*	*High*	*Low*	*High*	*Two*	*Three*	*Two*	*Three*	*No*	*Yes*	*Mean*
π_1	193.22	181.89	164.9	191.01	169	177.56	182.45	179.69	180.32	190.07	169.94	193.08	166.93	180.01
π_2	190.5	170.38	178.08	180.11	179.2	151.26	208.05		179.65	185.41	173.9	194.56	164.75	179.65
π_3	106.59	101.81	98.88	106.56	98.3	96.62	108.24	77.01	127.85	107.5	97.36	115.58	89.28	102.43
α_{A1}	244.83	209.46	177.03	224.23	196.65	226.45	194.43	196.65	224.23	262.39	158.49	234.72	186.16	210.44
α_{A2}	259.7	266.28	247.24	278.11	237.37	211.86	303.62		257.74	379.49	135.99	262.41	253.07	257.74
α_{A3}	116.04	96.43	87.88	117.94	82.29	62.51	137.73	61.84	138.4	128.04	72.19	106.17	94.07	100.12
α_{B1}	265.52	303.05	350.38	326.47	286.16	306.31	306.32	146.99	465.64		306.32	462.03	150.6	306.32
α_{B2}	283.47	289.77	301.23	309.51	273.48	288.46	294.52		291.49		291.49	281.65	301.33	291.49
α_{B3}	96.76	94.78	97.89	115.57	77.38	69.23	123.72	14.07	178.88		96.48	105.81	87.14	96.48
β_1	233.14	202.28	194.45	259.65	160.27	222.72	197.19	164.37	255.55	271.3	148.61	265.53	154.38	209.96
β_2	104.83	98.63	116.48	119.8	93.5	99.85	113.45		106.65	143.68	69.61	107.15	106.14	106.65
β_3	47.62	42.13	45.59	49.31	40.92	33.94	56.3	29.06	61.17	52.65	37.59	46.16	44.07	45.12
$\beta_{dis,1}$	106.26	129.06	144.49	135.29	117.91	143.73	109.47	106.27	146.94	160.81	92.39		126.6	126.6
$\beta_{dis,2}$	146.07	159.68	150.38	177.01	127.08	148.24	155.84		152.04	242.38	61.71		152.04	152.04
$\beta_{dis,3}$	49.29	46.76	53.81	53.67	46.24	38.85	61.06	17.79	82.11	73.84	26.07		49.95	49.95
γ_1	222.85	181.47	146.82	190.22	177.21	178.8	188.62	172.04	195.38	191.31	176.11	199.4	168.02	183.71
γ_2	739.69	841.72	752.8	827.81	728.33	705.42	850.72		778.07	792.39	763.75	861.76	694.38	778.07
γ_3	424.82	414.53	401.74	429.09	398.29	423.95	403.43	283.75	543.63	449.36	378.02	475.9	351.49	413.69
δ_1	161.25	172.89	194.45	197.46	154.93	199.74	152.65	158.92	193.47	197.66	154.73	211.07	141.32	176.2
δ_2	170.62	198.5	170.11	193.51	165.98	212.55	146.94		179.75	253.59	105.9	222.93	136.56	179.75
δ_3	59.7	54.29	53.27	57.32	54.18	63.46	48.04	33.16	78.34	66.99	44.51	63.29	48.21	55.75
Mean	193.94	189.61	182.24	203.97	173.21	184.3	192.89	124.92	231.04	216.18	165.25	224.7	158.04	

small. When the sample sizes are large, this problem goes away: Variations in observed share due to sampling error are minimized, and the discrepancies in observed market share are attributed primarily to heterogeneity. We believe that it is these fundamental information characteristics of aggregate data that explain our findings that segmentation recovery is difficult unless the sample size in the data is large.

The logic in the preceding paragraph suggests that the problem of recoverability is exacerbated when the number of latent segments being estimated is large. The only information we have to estimate a two-segment latent-class model is the discrepancies in observed aggregate share movements with the expected movements from the homogeneous model. To estimate a three-segment model, the only information available is in the discrepancies in observed aggregate share movements with the expected movements from the two-segment model. Similarly, to estimate a four-segment model, the only information available is in the discrepancies in observed aggregate share movements with the expected movements from the three-segment model, and so on. Because the marginal discrepancies typically become rapidly smaller as the number of segments increases, the exploitable information in the aggregate data

also decreases, and we would expect the difficulty of model estimation to increase rapidly.

Our theoretical results for model identification presume that the analyst knows the true number of segments. In real data situations, this is rarely the case, and so the analyst also needs to estimate the number of segments using model selection measures such as the BIC score. The arguments we have made about the diminishing information available for the estimation of increasing numbers of segments from aggregate data suggest that estimation of the number of segments is also difficult with aggregate data, particularly when the number of segments is large. To investigate the difficulty of estimating the number of segments, we performed a simple experiment with data from our main simulation experiment for the following cell (see Table 1): PRICEVAR = low, DISPLAY = yes, HETERO = medium, DELTA = low, NUMBRND = 3, NUMSEGS = 3. As in the main experiment, we considered several distinct values for the sample size n. For each sample size, we generated 25 realizations of the choice data. For each data realization, we estimated the number of segments S using the BIC score. The only values of S we entertained were 1, 2, 3, and 4. Bear in mind that, because of the cell we chose to study, the true value of S is 3. We counted the number of times (of 25) that the BIC score resulted in S being estimated as 1, 2, 3, or 4. Table 7 summarizes the results. The numbers in each row reveal the sample distribution for the estimate of S for each value of n. The results are entirely consistent with what might be expected from our previous arguments. The household estimates have access to richer data and, in this case, arrive at the correct value of S every time. Estimation of latent classes in store data is based on the exploitation of discrepancies of market shares from simpler models. Because the discrepancies are more difficult to observe for smaller sample sizes, as we mentioned previously, store data estimates of S tend to favor lower values. In larger samples, the discrepancies are easier to observe, and correct estimation of S also becomes easier. In Table 7, for smaller sample sizes, we observe that the dominant estimate for S from store data is S = 2, which is smaller than the true number, 3. The error is corrected as sample size increases. Note that at the smallest sample size, in most of the data realizations, the aggregate market shares do not exhibit departures from the homogeneous logit model that are sufficiently large to reject the no-heterogeneity model in favor of the S = 2 model.

The information characteristics we discussed previously are inherent to the data and affect all estimation procedures. In this article, we presented results only from maximum likelihood estimation with direct quasi-Newton maximization of the log-likelihood function. We also tried other methods on a limited scale: (1) maximization of the log-likelihood function with simulated annealing, (2) indirect maximum likelihood estimation with data augmentation using the expectation maximization algorithm, (3) Bayesian estimation of the posterior mean with Markov chain Monte Carlo methods without data augmentation on segment membership of individual transactions, and (4) Bayesian estimation of the posterior mean with Markov chain Monte Carlo methods with data augmentation. With small sample sizes in simulated data sets that are similar to the ones we discussed in the previous section, none of the methods offered successful recovery of segmentation structure. With large sample sizes, all the methods recovered the latent-class parameters correctly. Thus, we conjecture that the difficulties in recovering segmentation structure when large sample sizes are not available may be primarily due to fundamental information limitations of aggregate data rather than to inadequacies of particular estimation algorithms.

An important question that requires detailed investigation in further research is the differential impact of model misspecification on household and store data estimates. We conjecture that misspecification of the household-level model exaggerates the appearance of heterogeneity across households in aggregate date. Also of interest is the impact of misspecification of the heterogeneity distribution itself. For example, there is considerable debate on the question of continuous versus discrete specifications of the heterogeneity distribution. How does misspecification in this domain affect bias in estimates from household versus store data?

Although our prognosis for the estimation of latent-class models with only store-level data is not too optimistic, we believe a fruitful area for further research is the development of models that combine household panel data and aggregate store-level data. Such a combination could overcome the representativeness problems of panel data and the large inaccuracy of estimates that are based on store data alone. Russell and Kamakura (1994) take an important first step in this direction of combining micro and macro data, and Chintagunta and Dubé (2003) propose an approach to estimate random coefficients brand choice models by accounting for potential endogeneity of prices and combining household panel and store data.

Table 7
DISTRIBUTION OF THE ESTIMATED NUMBER OF SEGMENTS IN 25 REPLICATIONS

	Estimation with Household Data				*Estimation with Store Data*			
Number of Households	*S = 1*	*S = 2*	*S = 3*	*S = 4*	*S = 1*	*S = 2*	*S = 3*	*S = 4*
350	0	0	25	0	13	12	0	0
590	0	0	25	0	0	25	0	0
1000	0	0	25	0	0	25	0	0
1680	0	0	25	0	0	25	0	0
2830	0	0	25	0	0	9	16	0
4760	0	0	25	0	0	1	24	0
8000	0	0	25	0	0	0	25	0
13,450	0	0	25	0	0	0	25	0

Notes: The true number of segments is 3.

REFERENCES

Andrews, Rick, Andrew Ainslie, and Imran Currim (2002), "An Empirical Comparison of Logit Choice Models with Discrete Versus Continuous Representations of Heterogeneity," *Journal of Marketing Research*, 39 (November), 479–86.

Ben-Akiva, Moshe and Steven Lerman (1985), *Discrete Choice Analysis*. Cambridge: Massachusetts Institute of Technology Press.

Berry, Steve, M. Carnal, and P.T. Spiller (1997), "Airline Hubs: Costs, Markups, and the Implications of Consumer Heterogeneity," working paper, Department of Economics, Yale University.

Besanko, David, Jean-Pierre Dubé, and Sachin Gupta (2003), "Competitive Price Discrimination in a Vertical Channel Using Aggregate Retail Data," *Management Science*, 49 (9), 1121–38.

Bucklin, Randolph E. and Sunil Gupta (1992), "Brand Choice, Purchase Incidence, and Segmentation: An Integrated Modeling Approach," *Journal of Marketing Research*, 29 (May), 201–215.

——— and ——— (1999), "Commercial Adoption of Advances in the Analysis of Scanner Data," *Marketing Science*, 18 (3), 247–73.

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

———, Gary J. Russell, and V. Srinivasan (1998), "A Relationship Between Market Share Elasticities and Brand Switching Probabilities," *Journal of Marketing Research*, 35 (February), 99–113.

Chiang, Jeongwen (1991), "A Simultaneous Approach to the Whether, What, and How Much to Buy Questions," *Marketing Science*, 10 (4), 297–306.

Chintagunta, Pradeep K. (1992), "Heterogeneity in Nested Logit Models: An Estimation Approach and Empirical Results," *International Journal of Research in Marketing*, 9 (2), 161–75.

——— (1993), "Investigating Purchase Incidence, Brand Choice, and Purchase Quantity Decisions of Households," *Marketing Science*, 12 (2), 184–208.

——— (2001), "Endogeneity and Heterogeneity in a Probit Demand Model: Estimation Using Aggregate Data," *Marketing Science*, 20 (4), 442–56.

——— and Jean-Pierre Dubé (2003), "Estimating an SKU-Level Brand Choice Model Combining Household Panel Data and Store Data," working paper, Graduate School of Business, University of Chicago.

———, 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.

Draganska, Michaela and Dipak Jain (2002), "Product Line Length Decisions in a Competitive Environment," working paper, Graduate School of Business, Stanford University.

Food Marketing Institute (2002a), *Food Retailing in the 21st Century*. Washington, DC: Food Marketing Institute.

——— (2002b), *Key Facts*. Washington, DC: Food Marketing Institute.

Gourieroux, Christian and Alain Monfort (1995), *Statistics and Econometric Models*, Vol. 1. New York: Cambridge University Press.

Gupta, Sachin and Pradeep K. Chintagunta (1994), "On Using Demographic Variables to Determine Segment Membership in Logit Mixture Models," *Journal of Marketing Research*, 31 (February), 128–36.

———, ———, Anil Kaul, and Dick 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.

Gupta, Sunil (1988), "Impact of Sales Promotion on When, What, and How Much to Buy," *Journal of Marketing Research*, 25 (November), 342–55.

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.

McFadden, Daniel and Kenneth Train (2000), "Mixed MNL Models for Discrete Response," *Journal of Applied Econometrics*, 15 (5), 447–70.

Pace, Luigi and Alessandra Salvan (1997), *Principles of Statistical Inference from a Neo-Fisherian Perspective*. Singapore: World Scientific.

Prevision Corporation (1995), *1994 Marketplace Data User Survey: Summary Results*. Wellesley, MA: Prevision Corporation.

Russell, Gary J. and Wagner A. Kamakura (1994), "Understanding Brand Competition Using Micro and Macro Scanner Data," *Journal of Marketing Research*, 31 (May), 289–303.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, 6 (4), 461–64.

Seetharaman, P.B. (2001), "Estimating Disaggregate Heterogeneity Distributions Using Aggregate Data: A Likelihood-Based Approach," working paper, Olin School of Business, Washington University at St. Louis.

———, Andrew Ainslie, and Pradeep K. Chintagunta (1999), "Investigating Household State Dependence Effects Across Categories," *Journal of Marketing Research*, 36 (November), 488–500.

Shenton, L.R. and K. Bowman (1977), *Maximum Likelihood Estimation in Small Samples*. New York: Macmillan.

Thompson, S.K. (1992), *Sampling*. New York: John Wiley & Sons.

Zenor, Michael J. and Rajendra Srivastava (1993), "Inferring Market Structure with Aggregate Data: A Latent Segment Logit Approach," *Journal of Marketing Research*, 30 (August), 369–79.