RICK L. ANDREWS, ASIM ANSARI, and IMRAN S. CURRIM*

A study conducted by Vriens, Wedel, and Wilms (1996) and published in *Journal of Marketing Research* found that finite mixture (FM) conjoint models had the best overall performance of nine conjoint segmentation methods in terms of fit, prediction, and parameter recovery. Since that study, hierarchical Bayes (HB) conjoint analysis methods have been proposed to estimate individual-level partworths and have received much attention in the marketing research literature. However, no study has compared the relative effectiveness of FM and HB conjoint analysis models in terms of fit, prediction, and parameter recovery. To conduct such a comparison, the authors employ the simulation methodology proposed by Vriens, Wedel, and Wilms with some modification. The authors estimate traditional individual-level conjoint models as well. The authors show that FM and HB models are equally effective in recovering individual-level parameters and predicting ratings of holdout profiles. Two surprising findings are that (1) HB performs well even when partworths come from a mixture of distributions and (2) FM produces good parameter estimates, even at the individual level. The authors show that both models are quite robust to violations of underlying assumptions and that traditional individual-level models overfit the data.

Hierarchical Bayes Versus Finite Mixture Conjoint Analysis Models: A Comparison of Fit, Prediction, and Partworth Recovery

Conjoint analysis is one of the most popular market research procedures for assessing how consumers with heterogeneous preferences trade off the various benefits they derive from product or service attributes (e.g., Wittink and Cattin 1989). Analysis of trade-offs driven by heterogeneous preferences provides critical input for many marketing decisions, such as new product design, positioning, and pricing. Prior research has developed several procedures for modeling heterogeneity in consumer preferences. The traditional two-stage conjoint analysis procedure involves (1) estimating individual-level partworth utilities for attribute levels and (2) if segmentation is of interest to the marketing manager, clustering the individual-level partworths to derive segment-level partworths (e.g., Currim 1981; Green and Krieger 1991).

In the past decade, many integrated conjoint analysis methods have emerged that simultaneously segment the market and estimate segment-level partworths. A simulation study by Vriens, Wedel, and Wilms (1996; VWW) compared nine metric conjoint analysis methods, some traditional two-stage procedures and some integrated. The study by VWW, which was conducted at the segment level and focused on segment-level partworths, found that finite mixture (FM) conjoint models performed best in terms of model fit, prediction, and parameter and segment recovery.

Since VWW's study, hierarchical Bayes (HB) conjoint analysis methods (e.g., Allenby and Ginter 1995; Lenk et al. 1996) have emerged with much potential for representing heterogeneity in consumer preferences. In HB conjoint analysis models, instead of a semiparametric approach to modeling heterogeneity, a continuous population distribution is assumed for modeling the variation in individual-level partworths. Evidence exists that HB methods can recover heterogeneity and estimate individual-level partworths, even when individual-level least squares estimators do not exist because of insufficient degrees of freedom.

*Rick L. Andrews is an associate professor, Department of Business Administration, University of Delaware (e-mail: andrewsr@udel.edu). Asim Ansari is an associate professor, Department of Marketing, Graduate School of Business, Columbia University (e-mail: maa48@columbia.edu). Imran S. Currim is Corporate Partners Research Scholar and Professor, Graduate School of Management, University of California, Irvine (e-mail: iscurrim@uci.edu). The authors contributed equally to the article and are listed in alphabetical order. The first author acknowledges the support of a research grant from the University of Delaware.

Some researchers have compared FM and HB conjoint analyses methods on real data sets (see Allenby and Ginter 1995). However, no study has comprehensively compared the relative effectiveness of FM and HB conjoint analysis models in terms of fit, recovery of individual-level partworths, and prediction in carefully controlled simulations. Therefore, it is not known how the FM and HB models compare in recovering heterogeneity under different situations.

In their *JMR* guest editorial, Carroll and Green (1995, p. 389) state, "New developments in conjoint analysis are arriving so fast that even specialists find it difficult to keep up. Hierarchical Bayes models, latent class choice modeling, and individualized hybrid models are only a few of the new approaches and techniques that are arriving on the research scene." Furthermore, they note that the gap between academics and practitioners has not narrowed appreciably: "Part of the problem is the lack of critical comparisons among competing techniques." They cite VWW as an example of such a comparison study, in which model and method comparisons are made at a synthetic data level. Finally, Carroll and Green (1995, p. 389) suggest that comparisons among competing analysis techniques be made by impartial researchers with no vested interests in the performance of a model: "Perhaps the Marketing Science Institute or an AMA task force could be used to initiate procedures by which researchers other than the model's own developers can compare the competing models."

In this vein, we use simulated (synthetic) metric conjoint data based largely on the design by VWW (with one nonsignificant factor removed and two new factors added) to compare the relative performance of HB and FM conjoint analysis models. The performance measures include fit, parameter recovery at the individual level, and predictive accuracy on holdout profiles. An essential advantage of synthetic data (e.g., over data from real-world conjoint studies) is that the true parameter values at the individual level are known so that the difference between actual and estimated parameters can be computed.

In some of our experimental conditions, each true partworth is normally distributed across consumers, and in other conditions, the distribution of a partworth is a mixture of normal distributions with two or three mixture components. The data-generation process, though likely realistic, does not exactly match the data structure described by FM models because we allow for within-component heterogeneity (typically, FM models do not describe within-component heterogeneity). Similarly, the data-generation process does not exactly match the structure described by a typical HB continuous heterogeneity model, because some preferences are obtained from mixtures of normal distributions. The vast majority of HB models and applications assume unimodal population distributions (e.g., the normal distribution) for modeling heterogeneity; for an exception, see Allenby, Arora, and Ginter (1998).[1] Therefore, an important contribution of this study is that it assesses the robustness of the models to violations of their underlying assumptions. As another test of robustness, preferences are distributed according to mixtures of gamma distributions in some data conditions, not the normal distributions assumed by the HB model. Information on the robustness of the models to violations of their underlying assumptions about modality and the shape of the distribution is important for marketing scientists in both academic and corporate settings.

In contrast to VWW's study, which focused on comparisons of segment-level partworths, our study focuses on individual-level partworths. Comparing preferences at the individual level is a severe test for the FM models, because the models are known for producing segment-level preference information, not individual-level information. The individual-level estimates produced by the FM model are formed using posterior segment probabilities to calculate a weighted average of segment-level preference estimates. These estimates have been criticized in the literature as being restrictive because individual-level estimates are constrained to lie in the convex hull of the segment-level estimates (see Allenby and Rossi 1999; Wedel et al. 1999). An important contribution of this study is to assess the accuracy of individual-level parameter estimates and predictive capabilities of FM versus HB across a variety of experimental conditions.

In their recent book on market segmentation, Wedel and Kamakura (2000, p. 327), in comparing discrete (FM) and continuous (HB) representations of heterogeneity, conclude that "both discrete and continuous representations therefore seem to have some (advantages and) disadvantages, and under which conditions one of the two is most appropriate remains an empirical question." In the next section, we describe the design of the Monte Carlo study intended to address this question, including the data, models, and performance measures. We then present the results of the study and discuss implications and conclusions.

DESIGN OF THE MONTE CARLO STUDY

Data

The reader is referred to VWW (p. 78) for the rationale regarding choice of factors and levels. Seven factors were experimentally manipulated for this study:

1. the number of mixture components (one, two, or three components),
2. the masses of the mixture components if more than one (equal, unequal),[2]
3. the separation of mixture components if more than one (similar, dissimilar),
4. the within-component distribution of heterogeneity (normal, gamma),[3]
5. the within-component variances of distributions (.05, .10),
6. the number of profiles (18, 27), and
7. the error variance (5%, 35%).

Factors 2 and 3 are not meaningful if there is only one mixture component, so the bulk of the analysis is based on data sets in which there are two or three mixture components, producing $2^7 = 128$ experimental conditions. Five models are estimated per experimental condition (described in the next section), so $128 \times 5 = 640$ observations are statistically

[1]We do not consider an HB model that incorporates mixtures of normal distributions because of the computational and inferential difficulties of dealing with the "label-switching" problem when the Gibbs sampler is used for inference (for an insightful discussion of this problem, see Celeux, Hurn, and Robert 2000). In any case, the vast majority of applications in practice and in the literature use unimodal HB specifications, so the comparison of FM and unimodal HB models is relevant.

[2]VWW assume equal masses of the mixture components.

[3]VWW assume all normal distributions for preferences.

analyzed. To make the computational burden manageable, there is one replication per experimental condition, as in VWW's study. The power of this design to detect medium-sized main effects (effects accounting for approximately 5.88% of the variance of the dependent variable) at a significance level of .05 is more than 99% (Cohen 1988). For an effect accounting for as little as 2.20% of the variance in the dependent variable, the power is still 96%. Even if the effects are small (say, accounting for 1% of the variance of the dependent variable), the power is still quite good (70%).

Each data set contains the evaluations of 150 consumers on either 18 or 27 profiles (Factor 6).[4] We take the conjoint designs containing the 18 or the 27 profiles from Hahn and Shapiro (1966). The conjoint design varies six product attributes at three levels each. Regardless of whether 18 or 27 profiles are used for parameter estimation, we generate consumer evaluations of 8 additional holdout profiles to assess the predictive validity of the test models.

The range for the number of profiles enables us to assess the empirical effect of increasing the number of observations or degrees of freedom on parameter estimates and prediction. We expect better performance as degrees of freedom increase. Regarding the expected effects of the other factors, a greater number of components may adversely affect the unimodal HB model, whereas the FM model is designed to handle more than one component. Larger error variances are expected to decrease the performance of all models. When masses of the mixture components are unequal, performance of the FM model could decline because smaller components are more difficult to identify, whereas greater separation among components is likely to improve the performance of the FM model compared with the others. Gamma distributions for heterogeneity may penalize the HB models (which assume normal distributions of heterogeneity) more than the FM models (which assume discrete distributions of heterogeneity). Larger within-component heterogeneity is expected to worsen the performance of the FM model (for details, see VWW). Although such expectations follow directly from statistical estimation theory, the extent to which the factor levels affect the relative performance of FM versus HB models on various performance measures is unknown and is an empirical question.

The generation of the conjoint data closely follows that by VWW, so we refer the reader to that study for details. Essentially, the true partworths for each component were randomly (uniformly) generated to be in the range of −1.7 to 1.7. We added within-component noise having either normal or gamma distribution (factor 4) to the component partworths to simulate within-component heterogeneity. The same shape parameter (1.7) was used for all the gamma distributions to keep the level of skewness constant. Smaller values of the shape parameter produce more extreme skewness, whereas larger values produce distributions that more closely resemble normal distributions. We generated the normal deviates to have mean zero and variance of .05 or .10 (Factor 5) before adding them to the true parameter values. We standardized the gamma deviates to have mean zero and variance of .05 or .10 before adding them to the true parameter values.

Given β_i, the partworths for subject i, observable utilities were computed as $U_i = X\beta_i$. Normally distributed error variances σ_ε^2 corresponding to 5% or 35% (see Factor 7) of the total variances ($\sigma_\varepsilon^2 + \sigma_U^2$) were added to U_i to obtain the profile ratings $Y_i = X\beta_i + \varepsilon_i$.

In most of the data sets, the partworths are generated from mixtures of distributions (see Factor 1). When there are two mixture components, two sets of partworths are generated as described previously, and consumers are randomly assigned to the two components, with either equal or unequal probabilities (see Factor 2). When there are two components and the masses of components are equal, each consumer has a 50% chance of being assigned to each component. When the masses of components are unequal, each consumer has a two-thirds chance of being assigned to one of the components and a one-third chance of being assigned to the other. When there are three components and the masses are equal, each consumer has a one-third chance of being assigned to each component. However, when the masses are unequal, 50% of consumers will be assigned to one component, and 25% will be assigned to each of the other two components.

Notice that in the conditions with two and three mixture components, it is possible that some partworths will not differ greatly across components because they are all generated randomly. Factor 3, the similarity of components, manipulates the similarity of components by multiplying all components by two in the dissimilar condition. This results in a greater separation of components if the error levels are the same (see VWW).

Figure 1 shows the histogram of one true coefficient chosen randomly from six experimental conditions. The panels on the left-hand side of the figure show coefficients with normal distributions, and the panels on the right-hand show coefficients with gamma distributions. Panel A shows a coefficient having a normal distribution with .05 variance. The coefficient in Panel B has the shape of a gamma distribution, again with variance .05. Panel C shows a coefficient with two normal components generated to be similar to each other (Factor 3) and to have equal masses (Factor 2). Panel D shows a similar scenario except that the coefficient is based on a mixture of gammas. Because the means of the three components in Panel E are not well separated (they are generated independently), it is not apparent from the figure that there are three components. Apparently, two of the three distributions overlap on the left-hand side of the panel. That some components may overlap for some coefficients implies that the tests on the number of components factor could be conservative. However, it is extremely unlikely that two or more components could overlap for all 13 parameters in a given data set, so the overlap is probably not a serious problem. Panel F shows a situation that clearly indicates a mixture of three gamma components that are roughly equal in size. Note that the scales of the coefficients in Panels E and F are larger than those in Panels C and D, because components in Panels E and F are intended to be more dissimilar (Factor 3).

Models

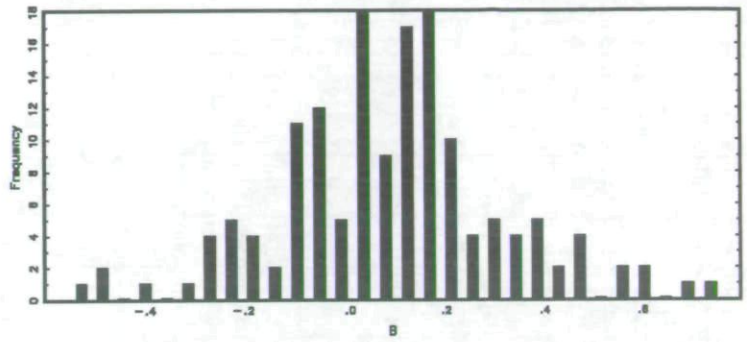
In addition to the two focal models, FM and HB, we also estimate individual-level conjoint models and an aggregate conjoint model for each data set.

[4]VWW vary the number of consumers between 100 and 200 but find mostly insignificant results for this factor, so we fix the number of consumers at 150.

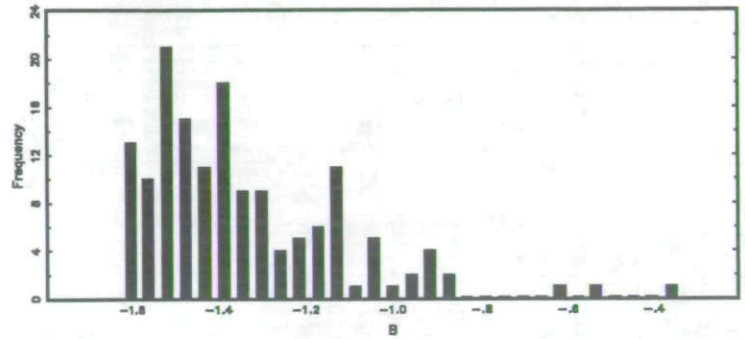
Figure 1

HISTOGRAM OF ONE TRUE COEFFICIENT FOR VARIOUS EXPERIMENTAL CONDITIONS

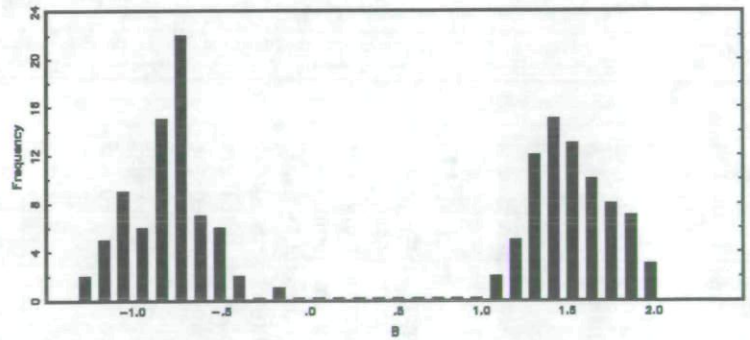
A: One Normal Component, .05 Within-Component Variance



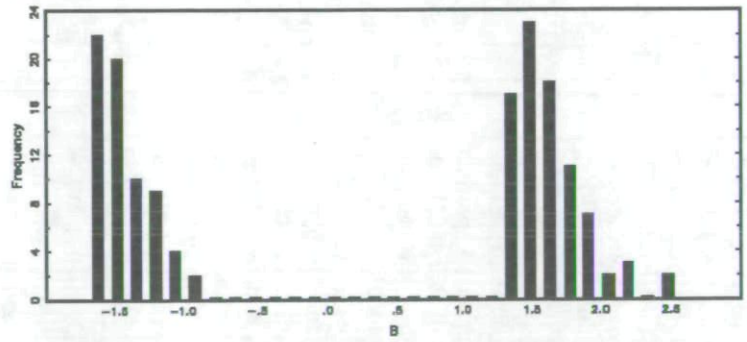
B: One Gamma Component, .05 Within-Component Variance



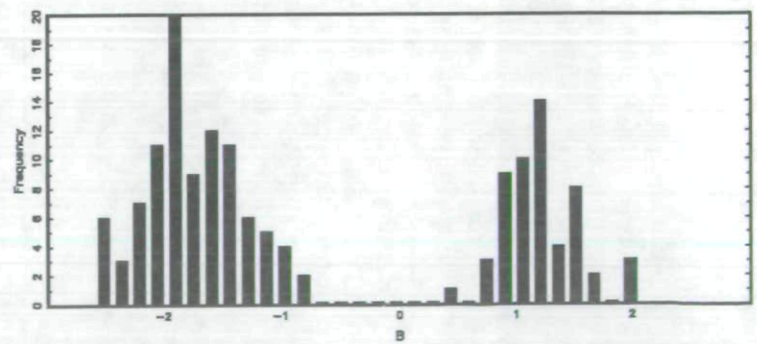
C: Two Similar Components, .05 Within-Component Variance, Equal Masses



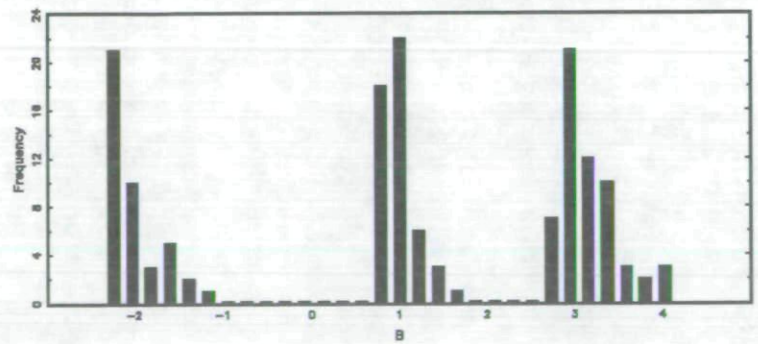
D: Two Similar Gamma Components, .05 Within-Component Variance, Equal Masses



E: Three Dissimilar Components, .10 Within-Component Variance, Equal Masses



F: Three Dissimilar Gamma Components, .10 Within-Component Variance, Equal Masses



FM models. The FM conjoint analysis model used in this study is based on that discussed by DeSarbo and colleagues (1992). Let

$i = 1, \ldots, I$ consumers;
$j = 1, \ldots, J$ choice alternatives;
$k = 1, \ldots, K$ derived components;
$l = 1, \ldots, L$ variables describing the alternatives;
Y_{ij} = the response to choice alternative j by consumer i;
Y_i = the $J \times 1$ column vector of responses by consumer i;
X_{jl} = the value of the lth variable for the jth alternative;
X_j = the $1 \times L$ row vector of variables for the jth alternative;
$X = [(X_{jl})]$, which is $J \times L$;
β_{lk} = the coefficient for the lth variable for the kth component;
β_k = the $L \times 1$ column vector of coefficients for the kth component;
$\beta = [(\beta_{lk})]$, which is $L \times K$;
Σ_k = a $J \times J$ covariance matrix estimated for component k; and
$\Sigma = (\Sigma_1, \Sigma_2, \ldots, \Sigma_k)$.

The density function for the response vector Y_i can be modeled as a mixture of distributions,

$$(1) \qquad H(Y_i; \alpha, X, \beta, \Sigma) = \sum_{k=1}^{K} \alpha_k g(Y_i | X, \beta_k, \Sigma_k),$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ are the mixing weights, interpreted as segment sizes, such that $0 < \alpha_k < 1$ and $\Sigma_k \alpha_k = 1$. When the preferences for alternatives Y_i are normally distributed ratings,

$$(2) \qquad g(Y_i | X, \beta_k, \Sigma_k) = (2\pi)^{-J/2} |\Sigma_k|^{-1/2}$$
$$\exp\left[-\frac{1}{2}(Y_i - X\beta_k)' \Sigma_k^{-1} (Y_i - X\beta_k)\right].$$

For the sake of parsimony and computational effort, we estimate models that constrain the variance of each alternative to be the same, for example, $\Sigma_k = \sigma^2_k I$.

For a sample of I consumers, the log-likelihood function is

$$(3) \qquad \ln L = \sum_{i=1}^{I} \ln\left[\sum_{k=1}^{K} \alpha_k g(Y_i | X, \beta_k, \Sigma_k)\right].$$

The parameters are estimated through maximization of Equation 3 by means of numerical optimization. The problem of local optima plagues FM estimation but not that of the other models included in this simulation. In cases in which the solution is obviously not correct (e.g., the likelihood is not at least as large as that of a model with fewer components, parameters are extremely large or unstable), we estimate the model again using a different set of starting values. The incidence of such local optima in our study appears to be small. In any case, diagnostic checking for actual applications might be more extensive, so the performance of FM in our simulation could be slightly conservative compared with actual applications.

We used the Bayesian information criterion (BIC; Schwarz 1978) and the consistent Akaike information criterion (CAIC; Bozdogan 1994) to determine how many components are appropriate for a given data set. The two criteria suggested the same number of components in all data conditions. The use of penalized likelihood measures for identifying the number of components K differs from the study by VWW, who assume that K is always equal to the true number of components. We estimated the FM conjoint models by increasing the number of components until the BIC was minimized. The results are given in Table 1. The BIC identifies the true number of components in 95 (4 + 42 + 49) of the 144 data sets. This should not necessarily be viewed as overfitting, because additional components might be required to accommodate the within-component heterogeneity in the data optimally.

Table 1
RESULTS OF FM CONJOINT MODEL ESTIMATION

	Number of Components Fitted							
True	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*Total*
1	4	5	5	1	1	0	0	16
2	0	42	14	5	2	0	1	64
3	0	0	49	11	1	2	1	64

Each additional component required 12 partworths (recall that the conjoint design varies six factors over three levels each, so $6[3 - 1] = 12$ partworths are required), an intercept, a variance, and a mixing weight. Therefore, models with 1, 2, 3, 4, 5, 6, and 7 components required 14, 29, 44, 59, 74, 89, and 104 parameters, respectively.

When the parameters were estimated and the number of components was determined, we estimated individual-level partworths using a weighted average (convex combination) of the segment-level partworths, in which the weights are the posterior probabilities of segment membership. The posterior probability for subject i of belonging to component k is computed as

$$(4) \qquad P(i \in k) = \frac{\hat{\alpha}_k g(Y_i | X, \beta_k, \Sigma_k)}{\sum_{k=1}^{K} \hat{\alpha}_k g(Y_i | X, \beta_k, \Sigma_k)}, k = 1, \ldots, K.$$

The estimates of individual-level partworths are then

$$(5) \qquad \hat{\beta}_i = \sum_{k=1}^{K} P(i \in k)\, \hat{\beta}_k.$$

We used individual-level partworths estimated in this manner in the computation of all measures of model performance, which we discuss subsequently in this section.

HB models. Instead of an FM specification, a continuous mixture model can be used to capture the heterogeneity in individual-level partworths. An HB approach can be used to specify such heterogeneity (see Allenby and Ginter 1995; Lenk et al. 1996). In the HB model, the sampling density for individual i can be written as

$$(6) \qquad f(Y_i; \beta_i, X, \sigma^2) = (2\pi\sigma^2)^{-J/2}$$
$$\exp\left[-\frac{\sigma^{-2}}{2}(Y_i - X\beta_i)'(Y_i - X\beta_i)\right],$$

where β_i is the vector of partworths for individual i, and σ^2 is the error variance. A continuous, unimodal population distri-

bution is then used to specify the heterogeneity across individuals. Typically, the multivariate normal distribution is used as a population distribution (see Allenby and Ginter 1995). In addition to the normal, $N(\mu, \Lambda)$, we also use the multivariate Student t distribution $t_v(\mu, \Lambda)$ with specified degrees of freedom v. The multivariate Student t distribution has fatter tails than the normal and therefore provides a robust alternative to the multivariate normal. Also, because the Student t distribution can be written as a scale mixture of the normal distribution, it adds little complexity when sampling-based Bayesian estimation methods are used for inference. The normal population distribution can be written as

$$(7)\quad g(\beta_i; \mu, \Lambda) = \sqrt{2\pi}|\Lambda|^{-1/2} \exp\left[-\frac{1}{2}(\beta_i - \mu)'\Lambda^{-1}(\beta_i - \mu)\right],$$

whereas the density for the Student t population distribution can be written as

$$(8)\quad g_t(\beta_i; \mu, \Lambda, \upsilon) \propto |\Lambda|^{-1/2} \left[1 + \frac{1}{\upsilon}(\beta_i - \mu)'\Lambda^{-1}(\beta_i - \mu)\right]^{-(\upsilon+L)/2}.$$

The moments of the t distribution are given as $E(\beta_i) = \mu$ if $\nu > 1$ and $\text{Var}(\beta_i) = \nu\Lambda/(\nu - 2)$ if $\nu > 2$.

The mean vector μ represents the mean partworths in the population, whereas the covariance matrix Λ for the normal distribution and the dispersion matrix Λ for the t distribution capture the extent of heterogeneity and the correlation in partworths across individuals. The lower the value of ν for the t distribution, the fatter its tails and therefore the more distinct the distribution is from the normal. In this study, we fitted models with $\nu = 4$ to guarantee sufficient difference from the normal distribution.

Hierarchical Bayes models also require priors over the hyperparameters μ and Λ and over σ^2 for inference. We used an inverse gamma, IG(a, b), prior for the residual variance, σ^2; a Wishart prior, $W[\rho, (\rho R)^{-1}]$, for the precision matrix, Λ^{-1}; and, finally, a multivariate normal prior, $N(\eta, C)$, for the population mean μ. Specifically, we set $a = 3$, $b = 1$ for the inverse gamma distribution (corresponding to a mean and variance of .5); $\rho = 14$ and $R = \text{Diag}(.1)$ for the Wishart distribution; and $\eta = 0$, $C = 10^3 I$ for the normal to obtain noninformative but proper priors. We performed inference using standard Gibbs sampling methods (for details on full-conditional distributions and choice of parameterizations, see Ansari, Essegaier, and Kohli 2000; Gelfand and Smith 1990). In each case, we ran the Gibbs sampler for 5000 iterations. We used a burn-in period of 2500 iterations, and therefore inferences are based on the last 2500 draws. We monitored the time-series plots of the Gibbs sampling draws to ensure convergence. The Gibbs sampling provides estimates for the hyperparameters, μ and Λ; the residual variance, σ^2; and the individual-level parameters, β_i. We used these individual-level coefficients in the computation of all performance measures that are discussed in the next section.

Individual-level models. Standard ordinary least squares regression analysis was used to estimate individual-level conjoint models. For each data set, 150 independent regressions were run. The partworths and fit measures were used in the computation of all measures of model performance.

Aggregate models. For each data set, we also ran a single regression analysis for all 150 subjects so that we could gauge the benefit of estimating individual-level parameters. In computing the model performance measures, all consumers have the same partworths.

Measures of Performance

To assess the performance of the models, we used four measures: one for fit, one for parameter recovery, and two for forecasting accuracy. Note that we could not compute other performance measures used by VWW because the analysis is performed at the individual level, not the segment level. We used point estimates of the individual-level parameters in the computation of all performance measures. For the HB models, we used the means of the posterior distributions as point estimates for individual-level parameters. Likewise, for the FM models, we used posterior probabilities as weights to form individual-level estimates from the segment-level estimates. The aggregate models assume the same responses for all individuals, and the individual-level models provide individual-level estimates directly. The performance measures are as follows

1. The percentage of variance explained by the conjoint models, R^2, is used as a measure of fit.
2. The measure of parameter recovery is the root mean square error between the true and estimated values of the partworths. The true partworths were saved in the process of data generation; they differ across consumers even within a component or segment because of the within-component variance (.05 or .10). The measure is computed as

$$(9)\quad \text{RMSE}(\beta) = \sqrt{\sum_{i=1}^{I}\sum_{l=1}^{L}\frac{\left(\hat{\beta}_{li} - \beta_{li}\right)^2}{LI}},$$

   where there are $L = 13$ predictors and $I = 150$ individuals.
3. As a measure of predictive accuracy, the root mean square error between the observed (Y_{ij}) and predicted ($\hat{Y}_{ij}$) preferences for the holdout sample is used:

$$(10)\quad \text{RMSE}(Y) = \sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(\hat{Y}_{ij} - Y_{ij}\right)^2}{IJ}},$$

   where there are $J = 8$ holdout profiles.
4. Following VWW, we used an alternative measure of forecasting accuracy, the percentage of first choice hits in the holdout sample (%1stCH). This represents the percentage of subjects for whom the highest preference among the holdout stimuli is predicted correctly. This measure is important because it is the choice rule most frequently applied in conjoint market simulations.

RESULTS OF THE MONTE CARLO STUDY

Table 2 shows the results of ANOVAs testing differences in the four response measures due to model type and the seven experimental factors. Because we estimated five models for each of 128 experimental conditions, the ANOVAs are based on 640 observations and 600 degrees of freedom for error. Similar to VWW, we include all main effects and all interactions involving model type. Model type is statistically significant at beyond .0001 for all four performance measures. Similarly, the error variance makes a significant difference for all four measures of performance, as does the model × error variance interaction.

The number of components in the data affects parameter recovery, fit, and prediction accuracy (RMSE[Y] but not

Table 2
F-TESTS OF MAIN AND INTERACTION EFFECTS ON PERFORMANCE MEASURES (n = 640; *p*-VALUES IN PARENTHESES)[a]

Source (Degrees of Freedom)	*F-Ratio RMSE(β)*	*F-Ratio R^2*	*F-Ratio RMSE(Y)*	*F-Ratio %1stCh*
Model (M)	1226.43	2884.77	169.42	66.01
(4)	(.0001)	(.0001)	(.0001)	(.0001)
Number of components	68.18	31.63	23.69	.42
(1)	(.0001)	(.0001)	(.0001)	(.5148)
Masses of components	8.99	4.46	1.02	18.12
(1)	(.0028)	(.0351)	(.3123)	(.0001)
Separation of components	853.21	13.31	605.49	.63
(1)	(.0001)	(.0003)	(.0001)	(.4293)
Within-components distribution	1.34	.37	.02	1.20
(1)	(.2478)	(.5427)	(.8979)	(.2735)
Within-components variance	7.41	.00	3.38	1.44
(1)	(.0067)	(.9493)	(.0665)	(.2302)
Number of profiles	12.09	14.06	1.22	3.44
(1)	(.0005)	(.0002)	(.2706)	(.0641)
Error variance	585.98	2507.61	915.92	492.98
(1)	(.0001)	(.0001)	(.0001)	(.0001)
Model × components	8.20	41.25	3.43	.30
(4)	(.0001)	(.0001)	(.0088)	(.8800)
Model × masses	5.01	5.67	.07	1.75
(4)	(.0006)	(.0002)	(.9904)	(.1381)
Model × separation	176.34	4.94	17.05	.65
(4)	(.0001)	(.0006)	(.0001)	(.6247)
Model × within-components distribution	.66	.06	.07	.31
(4)	(.6222)	(.9942)	(.9909)	(.8681)
Model × within-components variance	5.20	1.51	.27	.06
(4)	(.0004)	(.1988)	(.8997)	(.9926)
Model × profiles	5.20	8.98	.25	.21
(4)	(.0004)	(.0001)	(.9096)	(.9348)
Model × error variance	221.29	92.33	23.92	10.50
(4)	(.0001)	(.0001)	(.0001)	(.0001)
R^2	.93	.96	.80	.58

[a]The ANOVAs are based only on data sets with two or three preference components, because some of the factors (e.g., separation of components) are not meaningful when there is only one component.

%1stCH). The masses of components factor affects parameter recovery, fit, and predictive accuracy (%1stCH only). The separation of components factor affects parameter recovery, fit, and prediction (RMSE[Y] only). Surprisingly, the within-components distributions of partworths do not affect any of the performance measures. The within-components variance of partworths affects parameter recovery and marginally affects prediction (RMSE[Y] only). The number of profiles used in estimation affects parameter recovery and fit but only marginally affects prediction (%1stCH only). Overall, the parameter recovery measure RMSE(β) is affected by the largest number of experimental factors, and the predictive measure %1stCH is least affected. It is perhaps not surprising that %1stCH is the least affected by the factors and has the lowest R^2 (58%), because it is a cruder measure than the others.

Table 3 shows the means for the parameter recovery measure RMSE(β), and Table 4 shows the means for fit (R^2) and predictive validity (RMSE[Y] and %1stCH) by model type and experimental condition. Examining the overall model means at the bottom of Tables 3 and 4, we find that the FM and HB[5] models have the most accurate parameter estimates (RMSE[β]) and the best predictive accuracy (according to RMSE[Y] and %1stCH), whereas the individual-level conjoint models fit the best according to R^2. There are no significant differences between FM and HB in terms of parameter recovery or prediction accuracy—only in fit.

When a multivariate analysis of variance is estimated, none of the results reported in the article (the F-tests and *p*-values in Table 2 or the post hoc comparisons of means in Tables 3 and 4) is affected, because standard statistical packages conduct these tests for each dependent variable separately.

Parameter Recovery

We first examine parameter recovery for the conjoint models in Table 3. We also report the standard deviations of the true β coefficients to facilitate the evaluation of the practical significance of differences between methods. For RMSE(β), FM and HB are significantly (statistically and substantively) better than the individual and aggregate models (see the overall means at the bottom of Table 3). Surprisingly, there are no significant differences between FM and HB when there are two preference components,[6] but FM recovers parameters significantly better than HB when there are three components (a difference of at least .0418 is required for the RMSE[β] means by experimental condition;

[5]The usage of "HB" in the following paragraphs refers generally to HB and HB-t, because the models have similar performance.

[6]Recall that the results for one-component data sets are not included in the ANOVAs and therefore the post hoc comparisons, because Factors 2 and 3 (masses and separation) do not apply in the case of one-component data.

Table 3
PARAMETER RECOVERY RMSE(β) MEANS, BY MODEL TYPE AND EXPERIMENTAL CONDITION

Factor	*FM*[b]	*HB*	*HB-t*	*Agg*	*Ind*	*S.D. of True β*[c]
Components[a]						
1	.2676	.2172	.2138	.2736	.4626	.2693
2	.2833	.2951*	.2930*	1.0418*	.7558*	1.0396
3	.3101	.3562	.3571	1.2267	.8129	1.2245
Masses						
Equal	.3014	.3276	.3277	1.1913	.7896	1.1891
Unequal	.2920	.3237	.3224	1.0773*	.7791	1.0750
Separation						
Similar	.2874	.2684*	.2671*	.7892*	.5570*	.7877
Dissimilar	.3061	.3829	.3830	1.4794	1.0117	1.4764
Within-components distribution						
Normal	.2952	.3276	.3274	1.1548	.7886	1.1525
Gamma	.2982	.3237	.3227	1.1138	.7801	1.1116
Within-components variance						
.05	.2597*	.2992*	.3016*	1.1321	.8086	1.1297
.10	.3337	.3521	.3485	1.1365	.7601*	1.1344
Profiles						
18	.2964	.3392	.3394	1.1317	.8422	1.1292
27	.2970	.3121	.3106	1.1369	.7265*	1.1350
Error Variance						
5%	.2790	.2368*	.2331*	1.1434	.3960*	1.1430
35%	.3144	.4145	.4170	1.1252	1.1727	1.1211
Overall Means[d]	.2967[1]	.3256[1]	.3250[1]	1.1343[3]	.7844[2]	1.1321

[a]The means for one-component preferences were not analyzed in the ANOVAs in Table 2 and therefore are not tested for significance in this table.

[b]For FM models, the number of components is determined by BIC; HB models have normal heterogeneity; HB-t models have t-heterogeneity. Agg = aggregate conjoint analysis models; Ind = individual-level conjoint analysis models.

[c]S.D. of True β is the standard deviation of the true β values from the actual data sets, which is provided to facilitate the evaluation of the practical significance of differences among models.

[d]Superscripts on overall means by model type indicate significant differences (according to the least significant difference rule) at the $p < .05$ level, with the superior mean having a value of 1. The root mean square error value used in this analysis, from the RMSE(β) ANOVA in Table 2, is .1207.

*Indicates that the difference between the two means, generated under the corresponding method and the levels of the corresponding design attribute, is significant at the $p < .05$ level (as indicated by the least significant difference rule).

a difference of at least .0296 is required for the overall RMSE[β] means at the bottom of Table 3). The model × components interaction effect we observed for RMSE(β) in Table 2 is significant, because the models respond in different ways to increases in the number of components. The FM models handle two components as well as three components, but the aggregate models produce much less accurate parameter estimates when there are three components.

It is interesting that both FM and HB models recover parameters well regardless of the number of components, despite being handicapped: The HB models tested here are not designed for multiple components, and the FM models are not designed for continuous distributions of heterogeneity within components. It is also surprising that the individual models have poor parameter estimates, given their impressive fit statistics (see Table 4). Taken together, the fit and parameter recovery measures indicate that the individual-level models are overfitting the data.

Only the aggregate model produces significantly better parameter estimates when the components are of unequal sizes (probably because less of a compromise in parameter estimates is needed to accommodate the smaller component with different preferences). For the other models, the masses of the components do not affect parameter recovery. Most models have significantly better parameter recovery when components are less separated (more similar). The one exception is the FM model, which produces equally good parameter estimates whether components are similar or dissimilar. The FM models are designed to handle components with dissimilar preferences. Indeed, the FM model parameter estimates are significantly more accurate than those of HB when the components are dissimilar.

The distribution of within-components preferences (normal or gamma) does not affect parameter recovery, even though HB and HB-t models explicitly assume normal and t distributions, respectively, for preferences. Regarding the variance of within-components preferences, the FM and HB models produce significantly better parameter estimates when within-component heterogeneity is .05 rather than .10. Within-component variance is irrelevant to individual-level models because a separate model is fit to each individual's ratings. Overall, there is a significant difference as to how the various models respond to the within-component heterogeneity, according to the model × within-component variance interaction we observed in Table 2 ($p = .0004$).

Only the individual-level models produce better parameter estimates with larger numbers of profiles (27 versus 18). Statistical theory might suggest that all the models would produce better parameter estimates with more observations, but apparently the sample size issue is more critical with individual-level models, because there are only 4 degrees of freedom in the 18-profile condition but 13 in the 27-profile condition. As we observed in Table 2, this effect results in a significant interaction ($p = .0004$).

Table 4

FIT (R^2) AND PREDICTION ACCURACY (RMSE[Y] AND %1stCH) MEANS, BY MODEL TYPE AND EXPERIMENTAL CONDITION

	R^2					RMSE(Y)						%1stCH				
Factor	*FM*[b]	*HB*	*HB-t*	*Agg*	*Ind*	*FM*	*HB*	*HB-t*	*Agg*	*Ind*	*S.D.(Y)*[c]	*FM*	*HB*	*HB-t*	*Agg*	*Ind*
Components[a]																
1	.7371	.8118	.8207	.6945	.9119	1.0202	.9301	.9281	1.0949	1.0692	1.0922	.6617	.6779	.6808	.6554	.6458
2	.7566	.8116	.8168	.3202*	.8934	1.5427	1.4906	1.4896	2.7551*	1.7422	2.7508	.5829	.5947	.5982	.3758	.5525
3	.7602	.8205	.8253	.1862	.8855	1.6647	1.6197	1.6191	3.2864	1.8845	3.2820	.5982	.6106	.6116	.3596	.5603
Masses																
Equal	.7558	.8170	.8221	.2284	.8921	1.6236	1.5703	1.5698	3.0657	1.8273	3.0617	.5701	.5862	.5894	.3152	.5433
Unequal	.7610	.8151	.8201	.2779*	.8867	1.5838	1.5400	1.5389	2.9758	1.7993	2.9711	.6110	.6191	.6204	.4202*	.5695
Separation																
Similar	.7364	.8165	.8227	.2366	.8866	1.2160*	1.1217*	1.1204*	2.1426*	1.2818*	2.1398	.5747	.5967	.5983	.3795	.5511
Dissimilar	.7803*	.8156	.8194	.2697*	.8922	1.9914	1.9886	1.9883	3.8989	2.3448	3.8930	.6065	.6086	.6115	.3559	.5617
Within-components distribution																
Normal	.7614	.8176	.8223	.2532	.8901	1.6044	1.5588	1.5583	2.9951	1.8168	3.0870	.6047	.6083	.6105	.3630	.5660
Gamma	.7553	.8145	.8199	.2531	.8887	1.6030	1.5515	1.5504	3.0464	1.8099	2.9458	.5765	.5970	.5993	.3724	.5468
Within-components variance																
.05	.7679*	.8166	.8214	.2477	.8851	1.6101	1.5870	1.5873	3.0917	1.8703	2.9747	.5985	.6074	.6118	.3786	.5592
.10	.7488	.8155	.8207	.2586	.8937	1.5973	1.5233	1.5214	2.9498	1.7563	3.0581	.5826	.5979	.5980	.3568	.5536
Profiles																
18	.7629	.8198	.8254	.2480	.9222*	1.5624	1.5281	1.5274	2.9792	1.8308	2.7568	.6067	.6148	.6170	.3778	.5574
27	.7539	.8123	.8167	.2583	.8566	1.6450	1.5822	1.5813	3.0623	1.7958	3.2760	.5745	.5905	.5928	.3576	.5554
Error variance																
5%	.9034*	.9552*	.9572*	.2946*	.9661*	.9487*	.8252*	.8234*	2.7578*	.9147*	2.9903	.7123*	.7402*	.7414*	.4241*	.7196*
35%	.6134	.6769	.6849	.2117	.8127	2.2586	2.2851	2.2853	3.2837	2.7119	3.0425	.4689	.4651	.4684	.3114	.3932
Overall means[d]	.7584[3]	.8160[2]	.8211[2]	.2532[4]	.8894[1]	1.6037[1]	1.5551[1]	1.5544[1]	3.0208[3]	1.8133[2]	3.0164	.5906[1,2]	.6027[1]	.6049[1]	.3677[3]	.5564[2]

[a]The means for one-component preferences were not analyzed in the ANOVAs in Table 2 and therefore are not tested for significance in this table.

[b]For FM models, the number of components is determined by BIC; HB models have normal heterogeneity; HB-t models have t-heterogeneity. Agg = aggregate conjoint analysis models; Ind = individual-level conjoint analysis models.

[c]S.D.(Y) is the standard deviation of the Y values from the actual data sets, which is provided to facilitate the evaluation of the practical significance of differences among models.

[d]Superscripts on overall means by model type indicate significant differences (according to the least significant difference rule) at the $p < .05$ level, with the superior mean having a value of 1. The root mean square error values used in this analysis, from the ANOVAs in Table 2, are .0544 (R^2), .5479 (RMSE[Y]), and .1402 (%1stCH).

*Indicates that the difference between the two means, generated under the corresponding method and the levels of the corresponding design attribute, is significant at the $p < .05$ level (as indicated by the least significant difference rule).

Parameter recovery is usually better when error variance is smaller, as would be expected. The FM models are an exception—parameter recovery is not significantly affected by the amount of error variance. The HB models are slightly (but significantly) better than the FM model when error variance is 5%, but the FM model is substantially and significantly better than the HB models when error variance is 35%. It is not clear why HB models do not perform as well when the error variance is 35%, as this factor has nothing to do with multiple components in the data. Spurious overfitting of the additional error variance by HB models does not seem to be indicated, because the R^2 measure for the HB models is .6769 in this condition (see Table 4), near the expected value of .65.

Overall, there are no significant differences among the FM, HB, and HB-t models in terms of parameter recovery. The FM models produce good individual-level parameter estimates, though the estimates are constrained to lie in the convex hull of segment-level estimates and consequently have restricted variance. We verified this restricted range empirically—averaged across data conditions, the individual-level FM estimates have a range of 2.0024, compared with 2.9106 for the HB models and 3.2382 for the true parameters. Individual-level models have over 2.5 times more error in parameter recovery than the best models, on average, and the aggregate-level models have almost 4 times more error in parameter recovery than the best models. The advantages of modeling individual heterogeneity are quite palpable.

Fit

The R^2 results in Table 4 show that the various models respond differently to multiple components in the data. The FM, HB, and individual-level models are not significantly affected by an increase in the number of components,[7] whereas the aggregate models fit much worse as the number of components increases. (For statistical significance at the .05 level, a difference of at least .0188 is required for the R^2 means by experimental condition; a difference of at least .0134 is required for the overall R^2 means at the bottom of Table 4.) The striking differences across models produce a significant model × component interaction for the R^2 measure ($p < .0001$).

The masses and separation of components appear to make little difference in fit for most models, though the FM model fits significantly better with dissimilar components than with similar components, and the aggregate model tends to fit better with unequal-sized and dissimilar components. Both the model × masses and model × separation interactions are significant in Table 2. The distribution of within-component heterogeneity does not affect the models in different ways—none of the models fits normally distributed preferences differently than gamma distributed preferences.

Within-components variance does not significantly affect R^2, nor does the model × within-components variance interaction (see Table 2). The fit of most models is not affected by whether there are 18 or 27 profiles. However, the individual-level models appear to fit much better (more than 6% better) with smaller sample sizes, which possibly indicates some overfitting because there are only four degrees of freedom for error for these models with 18 profiles.

As mentioned, the effect of error variance is large. Note that with 5% error variance, we would expect R^2 to be approximately 95%; likewise, we would expect R^2 to be approximately 65% for the 35% error variance condition. Both FM and especially HB are quite close to these figures. The individual-level models appear to overfit significantly when error variance is 35%, because the average R^2 is 81%. All models fit significantly better when error variance is 5% rather than 35%.

Overall, there are significant differences in fit among the model types, and the best to worst models are individual-level, HB and HB-t, FM, and aggregate. However, the individual-level models did not perform nearly as well in terms of parameter recovery, as we showed in the previous section.

Prediction Accuracy

In our discussion of prediction accuracy, we focus primarily on the RMSE(Y) measure because it is a more exact measure than %1stCH. Note that we have also provided in Table 4 the standard deviations of the actual Y values from the validation samples to facilitate evaluation of the practical significance of the differences between methods.

Overall, the HB (and HB-t) and FM models have the best prediction accuracy, significantly better than the individual-level and aggregate models. Likewise, for %1stCH, HB, HB-t, and FM are not significantly different.

It appears that the error variance and the separation of components are the most powerful determinants of prediction accuracy. All models predict much better when error variance is 5% rather than 35%. Although HB appears to forecast better than FM when error variance is 5%, and FM appears to forecast better than HB when error variance is 35%, these differences are not statistically significant (a difference of at least .1898 is required for the RMSE[Y] means by experimental condition; a difference of at least .1345 is required for the overall RMSE[Y] means at the bottom of Table 4).

The prediction accuracy of the aggregate model falters significantly as the number of components increases, but this is not so for any of the other models. This is probably the primary explanation for the significant model × component interaction ($p = .0088$). The HB and FM models have equally good prediction accuracy even when there are two or three components.

Although the masses of components do not significantly affect prediction accuracy, the separation of components has a large effect on prediction accuracy (RMSE[Y]). All models predict considerably better when the components are similar (i.e., close together) rather than dissimilar (i.e., well separated). However, this finding could be partially an artifact of the way dissimilarity was created—in dissimilar conditions, the initial partworths (which were generated randomly) are multiplied by two. Larger partworths produce more variance in utilities σ_U^2 across consumers, which results in more error variance being added to the data (recall that error variance σ_ε^2 is set at either 5% or 35% of the total variance $[\sigma_\varepsilon^2 + \sigma_U^2]$), which could make parameter recovery and prediction more difficult. Similar findings with respect to the separation factor were reported by VWW. Therefore, we are cautious in interpreting this finding.

[7]See n. 6.

One interesting finding is that FM has better predictive validity than individual-level models, whereas VWW found no difference. An important difference between this study and VWW's is that VWW assume that the number of components for the FM model is fixed at the true number, whereas we use BIC to determine the number of components for each data set. As shown previously, the presence of within-component heterogeneity often results in fitting more than the true number of components. It is not unlikely that the larger number of components (compared with the true number) used in this study better accommodates the within-component heterogeneity and therefore leads to better predictive validity.

As with the parameter recovery measure RMSE(β), we observe that with regard to overall prediction accuracy (RMSE[Y]), there are no significant differences between the FM and HB models. For %1stCH, the same findings hold, except that individual-level models predict as well as FM models, in a statistical sense.

DISCUSSION AND CONCLUSIONS

This study compares HB conjoint models with unimodal population distributions to established methods such as FM conjoint models and individual-level models in a simulation experiment. Whereas VWW conducted the analysis at the segment level, our study is conducted at the level of the individual, which is a more demanding test of the FM models. Conducting the analysis at the segment level would probably be a more difficult test for HB models, because some clustering algorithm or an FM formulation would be needed to form the segments from the individual-level estimates. Segment-level partworths are useful for market segmentation, market summarization, and generation of new product or service ideas. Individual-level partworths are useful for mass customization and simulation of market outcomes for alternative offerings.

Despite receiving much attention in the recent marketing research literature, the performance of HB models has not been compared with that of established methodologies such as FM models in simulated settings. A key advantage of a simulated setting over a real-world conjoint data set is that in the simulated setting, the true parameter values are known, and in the real-world setting, these are unknown.

We summarize the major findings from the simulation experiment as follows:

1. Individual-level conjoint models fit the data well but produce poor parameter estimates and forecasts, which is indicative of overfitting.
2. Aggregate models are not the answer, because they perform significantly worse than all other models on all four performance measures.
3. The HB (both normal and t-heterogeneity) and FM models perform significantly better than individual-level and aggregate models in terms of partworth recovery and predictive accuracy, and they also fit better than the aggregate models. For three of the four performance measures, there are no significant differences overall between HB and FM—only in terms of fit (R^2) does HB perform significantly better than FM. Unlike individual-level models, both FM and HB use a combination of individual- and aggregate-level information to form individual-level partworth estimates.
4. The HB models produce partworths as accurate as those from FM models when there are only two components in the data, but the FM partworths are significantly more accurate when there are three components in the data. The surprising finding is that HB performs as well as it does when partworths come from a mixture of normal distributions with two or three components. This is impressive given that the model, as implemented in this study, is intended to fit unimodal distributions. Given the difficulties associated with the "label-switching" problem with multimodal HB models, the need for such elaborate models could be questioned.
5. Another surprising finding is that the FM models produce good individual-level parameter estimates. Given the discussion in the literature about how the individual-level estimates produced by an FM model are constrained to lie in the convex hull of segment-level estimates and therefore have restricted variance, we expected parameter recovery to be FM's weakest feature. Instead, parameter recovery was one of FM's strongest features.
6. The HB models do not recover partworths as well as FM models when error variances are large. The HB partworth estimates have significantly higher estimation error than those from FM when error variance is 35%. In contrast, the parameter recovery of the FM models is not significantly affected by the amount of error variance in the data. It is not clear why the performance of the HB models suffers when error variances are large. It does not appear that HB overfits the additional error variance as heterogeneity.
7. The FM parameter estimates are significantly more accurate than those of HB when components are dissimilar. The HB model probably becomes a less reasonable approximation to a mixture of normal components as the mixture components become farther apart. Alternatively, as explained previously, dissimilar components indirectly produce an increase in error variance, which seems to penalize the HB models (see the preceding point).
8. Overall, the FM and HB models are quite robust to violations of underlying assumptions, including the existence of within-component heterogeneity (FM), the distribution of within-component heterogeneity (HB), and multimodality (HB).

These findings have significant implications for researchers seeking to model heterogeneous preferences in conjoint analysis experiments. In their study of the commercial use of conjoint analysis, Wittink and Cattin (1989) find that the majority of commercial applications (54% of 1062 projects identified) used least squares to estimate partworth utilities. They report a median of 16 judgments per respondent for the typical application. In discussing these results, Wittink and Cattin (1989, p. 94) express concern: "Indeed, 16 judgments seem inadequate for the estimation of all parameters in a study using eight attributes and three levels per attribute in a partworth model." Our study validates their concern in that traditional individual-level conjoint models employing least squares as the estimation technique are shown to produce poor parameter estimates for partworth utilities and forecasts on a validation sample, though they fit the data well. These numbers indicate that there is substantial potential for improvement by using FM or HB models.

Improved parameter estimates and forecasts from FM and HB methods could also improve identification of the "optimal" product for market share or profit maximization (e.g., Green, Carroll, and Goldberg 1981) and optimized design for product line selection (e.g., Green and Krieger 1985). They could also improve the performance of models in which self-explicated attribute-level importances are obtained, in addition to the traditional evaluations of full

profiles (e.g., Green 1984). The likely extent of such improvements is an empirical question that needs to be addressed in further research. Other topics for further research are the performance of the models when there are even fewer observations per respondent (producing insufficient degrees of freedom for individual least squares estimates) and the performance of an HB specification that is capable of handling mixtures of normal distributions.

Likewise, the substantive implications of a wider range of parameter estimates produced by HB models compared with FM models need to be studied more extensively. The range of the estimates becomes important under a loss function that is nominal (0/1) rather than squared error. The nominal loss function is encountered when the number of consumers with parameters in a particular range (e.g., $\beta > -1$) is an important construct, such as in assessments of the fraction of the population with inelastic price elasticity. The range of heterogeneity also affects issues such as the optimal breadth of the product line.[8]

Finally, Wittink and Cattin (1989, p. 92) report that conjoint analysis is used by managers for a variety of purposes, such as new product/concept identification, competitive analysis, pricing, market segmentation, repositioning, advertising, and distribution. Therefore, improvements in parameter estimates and forecasts from FM and HB methods could have a relatively widespread impact on marketing practice.

REFERENCES

Allenby, Greg M., Neeraj Arora, and James L. Ginter (1998), "On the Heterogeneity of Demand," *Journal of Marketing Research*, 35 (August), 384–89.

——— and James L. Ginter (1995), "Using Extremes to Design Products and Segment Markets," *Journal of Marketing Research*, 32 (November), 392–403.

——— and Peter E. Rossi (1999), "Marketing Models of Consumer Heterogeneity," *Journal of Econometrics*, 89 (March/April), 57–78.

Ansari, Asim, Skander Essegaier, and Rajeev Kohli (2000), "Internet Recommendation Systems," *Journal of Marketing Research*, 37 (August), 363–75.

Bozdogan, H. (1994), "Mixture Model Cluster Analysis Using Model Selection Criteria and a New Informational Measure of Complexity," in *Multivariate Statistical Modelling*, Vol. 2, H. Bozdogan, ed. Dordrecht, The Netherlands: Kluwer Academic Publishers, 69–13.

Carroll, J. Douglas and Paul E. Green (1995), "Psychometric Methods in Marketing Research: Part I, Conjoint Analysis," *Journal of Marketing Research*, 32 (November), 385–91.

Celeux, Giles, Merrilee Hurn, and Christian P. Robert (2000), "Computational and Inferential Difficulties with Mixture Posterior Distribution," *Journal of the American Statistical Association*, 95 (September), 957–70.

Cohen, Jacob (1988), *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Currim, Imran S. (1981), "Using Segmentation Approaches for Better Prediction and Understanding from Consumer Mode Choice Models," *Journal of Marketing Research*, 18 (August), 301–309.

DeSarbo, Wayne S., Michel Wedel, Marco Vriens, and Venkatram Ramaswamy (1992), "Latent Class Metric Conjoint Analysis," *Marketing Letters*, 3 (3), 273–88.

Gelfand, Alan E. and Adrian F.M. Smith (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85 (June), 398–409.

Green, Paul E. (1984), "Hybrid Models for Conjoint Analysis: An Expository Review," *Journal of Marketing Research*, 21 (May), 155–69.

———, J. Douglas Carroll, and Stephen M. Goldberg (1981), "A General Approach to Product Design Optimization via Conjoint Analysis," *Journal of Marketing*, 45 (Summer), 17–37.

——— and Abba M. Krieger (1985), "Models and Heuristics for Product Line Selection," *Marketing Science*, 4 (1), 1–19.

——— and ——— (1991), "Segmenting Markets with Conjoint Analysis," *Journal of Marketing*, 55 (October), 20–31.

Hahn, G.J. and S.S. Shapiro (1966), *Statistical Designs of Experiments*. Schenectady, NY: General Electric Technical Information Series, Corporate Research and Development.

Lenk, Peter J., Wayne S. DeSarbo, Paul E. Green, and Martin R. Young (1996), "Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs," *Marketing Science*, 15 (2), 173–91.

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

Vriens, Marco, Michel Wedel, and Tom Wilms (1996), "Metric Conjoint Segmentation Methods: A Monte Carlo Comparison," *Journal of Marketing Research*, 33 (February), 73–85.

Wedel, Michel and Wagner A. Kamakura (2000), *Market Segmentation: Conceptual and Methodological Foundations*, 2d ed. Boston: Kluwer Academic Publishers.

———, ———, Neeraj Arora, Albert Bemmaor, Jeongwen Chiang, Terry Elrod, Rich Johnson, Peter Lenk, Scott Neslin, and Carsten Stig Poulsen (1999), "Discrete and Continuous Representations of Unobserved Heterogeneity in Choice Modeling," *Marketing Letters*, 10 (3), 219–32.

Wittink, Dick and Philipe Cattin (1989), "Commercial Use of Conjoint Analysis: An Update," *Journal of Marketing*, 53 (July), 91–96.

[8]We thank a reviewer for pointing this out to us.