# Product Launches with New Attributes: A Hybrid Conjoint-Consumer Panel Technique for Estimating Demand

Paul B. Ellickson, Mitchell J. Lovett, and Bhoomija Ranjan*

July, 2017

**Abstract**

We propose and empirically evaluate a new hybrid estimation approach that integrates choice-based conjoint with repeated purchase data for a dense consumer panel, and show that it increases the accuracy of conjoint predictions for actual purchases observed months later. Our key innovation lies in combining conjoint data with a long and detailed panel of actual choices, both before and after the product line introduction, for both survey respondents and a random sample of the target population. By linking the actual purchase and conjoint data, we can estimate preferences for attributes not yet present in the marketplace, while also addressing many of the key limitations of conjoint analysis, including sample selection and contextual differences. A counterfactual product and pricing exercise illustrates its managerial relevance.

**Keywords:** Conjoint, Revealed Preference, Stated Preference, Data Fusion, Predictive Validity, Choice Models.

*Paul and Mitchell are, respectively, the Michael and Diane Jones Professor of Marketing and Economics at the Simon Business School and an Associate Professor of Marketing, University of Rochester, 500 Wilson Blvd., Rochester, NY 14627. Bhoomija is an Assistant Professor at the International Business School, Brandeis University, 415 South Street MC032, Waltham, MA 02453. Paul's phone and email address are 585-273-1491 and paul.ellickson@simon.rochester.edu. Mitchell's phone and email are 585-276-4020 and mitch.lovett@simon.rochester.edu. Bhoomija's email is branjan@brandeis.edu. The authors want to thank participants of the Simon Marketing Brown Bag for their thoughts.

## *INTRODUCTION*

Forecasting new product sales to optimize a product's feature set and price point is a critical marketing research goal. However, when a product introduction involves attributes that are new to the marketplace, sales forecasts cannot be directly inferred from past purchases. The canonical solution to this problem is a class of conjoint analysis techniques including choice-based conjoint (Green and Rao 1971; Green and Srinivasan 1990; Wittink and Cattin 1989), in which a sample of existing consumers are repeatedly tasked with choosing one of several hypothetical products featuring potentially desirable combinations of characteristics (Ben-Akiva et al., 2015). This conjoint choice data is then used to estimate preferences and, ultimately, inform decisions related to the new products' attributes and pricing.

However, it is now well-known that conjoint can produce biased preference estimates (Ben-Akiva and Morikawa 1990; Horsky and Nelson 1992; Brownstone et al. 2000; Swait and Andrews 2003; Brooks and Lusk 2010) due to sample selection or contextual differences (e.g., Allenby et al. 2005) that can lead to preference distortions and inaccurate choice predictions (Ozer and Kamakura 2007 and Swait and Andrews 2003). In 2000, Dick Wittink wrote, "unfortunately, there is not much hard evidence that future market outcomes are predictable [by conjoint]" (Wittink, 2000), echoing previous (Orme et al., 1997) and subsequent (Wittink, 2003) calls for such investigations. The evidence on the predictive accuracy of conjoint for forecasting future choices remains surprisingly limited, leading most practitioners to refer to market share simulations as "preference shares", a tacit acknowledgement of their poor forecasting abilities (Orme, 2014). Perhaps the most promising avenue for improving the predictive power of conjoint lies in a growing set of hybrid data-fusion techniques that combine conjoint with actual purchase data (Swait and Andrews 2003; Allenby et al. 2005; Feit et al. 2010). By anchoring stated preferences to actual market choices, such data fusion approaches can leverage the strengths of both methods. This paper is aimed at furthering these goals.

To do so, we develop a new hybrid approach to data fusion that integrates choice-based conjoint with shopper panel repeated purchase data, and show that it increases the accuracy of conjoint predictions for actual purchases of new products observed months later. Our key innovation lies in combining conjoint data with a long and detailed panel of actual choices, both before and after the introduction of a new line of consumer products, for both survey respondents and a random sample of the target population. By linking actual purchase with conjoint data, we can recover preferences for attributes not yet present in the marketplace, while also addressing many of the key limitations of conjoint analysis, including sample selection and contextual differences.

We apply our method to a field study of the Greek yogurt category for a retailer that launched a new private label yogurt following a related conjoint study. We demonstrate that our hybrid approach can link conjoint data with actual purchases despite the clear presence of preference distortions and selection effects. We characterize the form and quantify the magnitude of these biases and show how our method accounts for them. In so doing, we demonstrate that our model predicts individual-level actual purchases better than both conjoint alone and a strong benchmark approach that combines key prior data-fusion techniques (Feit et al. 2010; Swait and Andrews 2003). Notably, we can accurately predict individual-level actual purchases months after the conjoint study, when the new product becomes available in stores. Moreover, we do so for a hold-out sample of individuals outside the estimation set. This challenging predictive task is both what is needed to make conjoint relevant to the managerial decisions it is tasked to support, and more involved than most extant evaluations. Finally, we demonstrate via counterfactual exercises that our method can inform decisions regarding optimal product selection and provide profit-enhancing store-level pricing recommendations for products with new to market features.

We make three main contributions to the current literature. First, we develop a novel hybrid approach to combining stated and revealed preference data for demand estimation that leverages the increasing ability to pair conjoint surveys with a rich and detailed panel of consumer choices. By fusing these data, we are able to mitigate many of the key biases that arise in conjoint while preserving the ability to forecast demand for products that are not yet present in the market. We are also able to identify the key drivers of bias and assess the ability of observed covariates to control for them. Second, we provide a new and detailed field evaluation of both conjoint and hybrid approaches that answers the call for greater demonstration of predictive validity.[1] Finally, we provide a detailed application showcasing the ability of these approaches to provide concrete managerial insights. We turn now to a more detailed discussion of the challenges of data fusion and our strategy for addressing these issues.

The existing literature has identified two key challenges to combining information from survey and actual purchase data (Adamowicz et al. 1994; Swait and Andrews 2003; Feit et al. 2010; Cherchi and de Dios Ortúzar 2002; Brownstone et al. 2000). First, conjoint respondents are often selected to have previous experience in the category. This can lead to biased coefficient estimates if not properly addressed (Manski and Lerman 1977; Feit et al. 2010).[2] Second, conjoint surveys can present a starkly different contextual environment than actual choices (Swait and Andrews 2003; Allenby et al. 2005), raising several thorny concerns. First, the choice setting may seem artificial (few options, not in store, unusual attributes), straining the respondent's ability to respond as they would to actual choices. Second, hypothetical choices may lead individuals to respond from a higher level of construal (Trope and Liberman, 2010), navigating trade-offs in a different manner than they would "in the wild" (Whitehead et al., 2008). Third, the survey setting itself may induce social desirability biases or other context-specific incentives leading survey respondents to choose options that they normally would not, such as supporting a

[1]This predictive validation is noteworthy, as we were unable to find any predictive validity tests for such difficult prediction tasks. In particular, most predictive validations are on hold-out tasks for the same individuals, rather than actual market choices (Wittink, 2000). Such predictions can establish internal, but not external, validity. Those on actual choices are scarce and limited to a single choice for the same (Toubia et al. 2003; Ozer and Kamakura 2007) or other individuals (Feit et al. 2010), or against aggregate market shares (Chang et al. 2009; Rogers and Renken 2003; Orme and Heft 1999; Feurstein et al. 1999). We could identify only two studies that examine predictions against repeated actual purchases at the individual-level, and both were from past purchases (Swait and Andrews 2003; Brooks and Lusk 2010). In contrast, our predictive validation incorporates (i) repeated individual-level purchases for (ii) both in-sample and a hold-out sample of individuals, and (iii) for multiple time periods, including months after the conjoint data was collected, which is the kind of timeline needed for supporting product launch decisions. Thus, our study provides a tighter demonstration of the accuracy of conjoint predictions as well as the predictive benefits of our approach to combining conjoint and actual purchases.

[2]Non-response that is correlated with preferences can also clearly be a problem. Our approach is to address this concern by conditioning on category-use variables (i.e. selection on observables approach).

public good, selecting "healthier" lifestyle choices (Whitehead et al., 2008), or exhibiting less price sensitivity than in actual choices due to lack of a clear budget constraint (Horsky and Nelson, 1992).

Our central modeling challenge lies in correcting for these preference distortions while still incorporating unique information from the conjoint data regarding new-to-market attributes. Building upon the earlier data-fusion frameworks of Swait and Andrews (2003) and Feit et al. (2010), our approach combines information from conjoint and actual purchase in a particular manner. Specifically, we use the trade-offs revealed in the actual purchase data to inform relative preferences for attributes already present in the market. We then use the conjoint choices to pin down the tastes for new attributes, while using attributes shared across both settings to carefully link choices through these relative tastes. Critically, we perform this fusion using only the stated preferences for the shared attributes that are believed to be free of distortions. We base this judgment on a mix of theory, prior research and empirical evidence, providing guidance for future improvements in practice.

We find that our model is able to identify and address several meaningful preference distortions in the conjoint data. We first estimate models using only conjoint and actual purchase data, and show that the (relative) trade-offs these models imply differ markedly for some attributes, but not others. One clear distortion relates to the estimated magnitude of the preference for inside goods versus the outside option. This distortion can arise both because of the selection of participants in the conjoint sample and from differences in conjoint and actual purchase contexts. We find that both sources of distortion exist in our empirical setting, highlighting the need to correct for each. The second set of distortions relates to weaker relative preferences in the survey data than the actual choice data. We find this distortion for price sensitivity and state dependence. The final distortion is that the relative preference for some attributes (e.g., fruit and fat content) are reversed in the conjoint data relative to actual choices. We demonstrate that our model can handle each of these distortions and more accurately represent the actual trade-offs, while still combining the data to produce a credible forecast for new product demand.[3]

Finally, we also are able to provide insight into the key drivers of bias in the conjoint analysis. Because we also observe actual purchases for the survey data sample (i.e. those who completed the conjoint), we can evaluate how and why their preferences differ from the full population. Specifically, we can evaluate the relative importance of selection (who takes part in the conjoint) versus contextual biases (how implied tastes differ across settings), and investigate how effective observable covariates are in mitigating the selection problem. To accomplish the latter, we draw upon information regarding consumer demographic and usage (RFM) variables. We find that the RFM variables address the selection problem better than demographic data alone, particularly as it relates to the pricing coefficient, which we find to be five times more biased without the RFM variables. This suggests that pricing elasticities may be understated in conjoint analysis as much due to selection as context effects. We find that the selection bias, after controlling for the RFM variables, is no more than one fourth the magnitude of the contextual bias. Simply including the RFM variables is not sufficient; handling the distortion in preferences that arise from conjoint elicitation is also critical for improved predictions. We conclude by performing a suite of counterfactuals that demonstrate the managerial insights from the improved predictive ability.

The rest of the paper is organized as follows. Section 2 describes the microeconomic and econometric framework for the model used. The description of the application context and data is presented in Section 3. Section 4 discusses the results including estimates, model comparison, predictive validity, and disentangling the nature of the conjoint biases. In Section 5, we use our model estimates to evaluate a managerial decision involving product optimization and store-level pricing. Finally, Section 6 concludes.

[3]We note that our conjoint application did not easily allow the kind of incentive compatibility practices recommended in the literature, e.g., Ding (2007) because the product was both novel and non-storable. In general, there are settings where incentive compatibility is not feasible for or desirable by the organization sponsoring the conjoint research (Dong et al., 2010). We follow the recommendations of Ding and Huber (2009) for cases when incentives are not possible, while noting that such improvements in conjoint design are a complement to the data fusion approach advocated here.

## *MODEL AND ESTIMATION*

We consider a firm choosing novel features to include in a new product launch and how to vary the product's price over different spatial locations (e.g. outlets or pricing zones). The retailer's decision problem is described in more detail in section 5. For now, we focus on a key informational input, namely the parameters of the underlying consumer demand system. This system characterizes the level of demand for products, the responsiveness to changes in the product characteristics, and how this responsiveness varies by spatial location. The central challenge for demand estimation, and the key motivation for combining conjoint data with actual purchase information, is that some product characteristics are not observed in the marketplace, particularly those associated with the new products under consideration.

In what follows, we first introduce a micro-founded model of consumer decision making. We then turn to the two datasets used in the estimation framework and explain how we link these datasets together. Finally, we present our estimation method and discuss our identification strategy. Throughout, we refer to units of analysis as individuals, but note that these could just as easily represent households or any other "micro-segments" classified a priori using demographic, location, and other observable variables.

### *Model Of Consumers*

Consumers indexed $i = 1, \ldots, n$ make choices over goods indexed by $j$ in a series of decision opportunities indexed by $t$. These decision opportunities may represent either actual purchases or tasks in a conjoint study. Goods are represented as preferences for bundles of attributes (i.e., characteristics space). The attributes are split into price, $p_{ijt}$, and a vector of other attributes or attribute levels, $X_{ijt}$. Individual $i$'s choices and indirect utilities at decision $t$ are related to a vector of preferences, $\beta_i$ and a price coefficient $\alpha_i$:

$$U_{ijt} = X_{ijt}\ \beta_i + p_{ijt}\alpha_i + \epsilon_{ijt}. \tag{1}$$

Assuming $\epsilon_{ijt}$ to be an unobserved iid Type I Extreme Value shock yields a logit formulation for the individual choice probabilities. Letting $C_{it}$ be the set of options available to consumer $i$ in decision $t$, the relevant choice probability is then

$$P_{ijt} = \frac{\exp(X_{ijt}\, \beta_i + p_{ijt}\alpha_i)}{\sum_{h=0}^{C_{it}} \exp(X_{iht}\, \beta_i + p_{iht}\alpha_i)}. \tag{2}$$

Consumers preferences are assumed to be a function of observed demographic variables, $Z_i$, and an unobserved dimension, $\nu_i$, assumed multivariate normal with mean zero and variance, $\Sigma$.[4] Thus,

$$(\beta_i, \alpha_i)' = \Delta_0 + Z_i\Delta + \nu_i. \tag{3}$$

*The Datasets*

Our primary interest is in the population distribution of $(\beta_i, \alpha_i)$, as indexed by the set of parameters $\Delta_0$, $\Delta$, and $\Sigma$. As noted above, the key estimation challenge stems from the fact that true 'purchase preferences' for a subset of the $K$ variables (product attributes) in $X_{ijt}$ cannot be inferred from actual purchase data alone. Of central interest are any new-to-market-attributes (NTMAs) that are not yet present in the market. This set is denoted $\mathcal{K}_{NTMA} \subset \mathcal{K}$ and includes $K_{NTMA}$ distinct attributes. For example, in our Greek yogurt application the private brand is new to the category, as are several new ingredients including fiber and Omega 3. Another subset, $\mathcal{K}_{Hidden}$, are available in the market, but the attractiveness of these $K_{Hidden}$ attribute levels cannot be identified from purchase data as these attribute levels are perfectly collinear with other variables in $X_{ijt}$ (see also Swait and Andrews (2003)). For example, if *all of exactly one brand's* products are organic, then purchase data cannot separately identify brand from organic preference.

To address this challenge, we incorporate data from a set of choice-based conjoint tasks and combine them with purchase information for the existing set of products. A conjoint experiment involves manipulating the set of product "profiles" that a consumer can choose in a series of "choice tasks". Because the set of attributes (and their levels) for each

[4]In principle, one could extend this formulation to include a semi-parametric distribution for the mixing distribution, such as a mixture of normal, to allow for even richer substitution patterns.

product is chosen by the researcher, these can include attributes not yet available in the marketplace, and they can be varied in such a way to allow the "hidden" attributes to be identified. Our estimation approach leverages both actual purchase 'revealed preference' (RP) data, as well as conjoint-elicited 'stated preference' (SP) data, to infer consumer preferences for the new and hidden attributes, while adjusting for selection and context bias issues raised by the conjoint experiment. We discuss these biases in detail shortly.

We draw these datasets from two different samples of individuals (or segments): a random sample 'rs' of consumers from the full population and a survey sample 'ss' that completes the conjoint tasks. Similar purchase data could be drawn from a retailer's loyalty card program, consumer panel data, or from online shopping histories. As is typical for conjoint analyses, the 'ss' sample was invited to complete the survey based on past purchases, meaning individuals without past purchases in the category are less likely to be included. By construction, data on conjoint choices are only available for the respondents in the survey sample. We now introduce additional notation to distinguish each group. The conjoint data are denoted $SP_{ss}$, indicating that they reflect *stated preference* outcomes for the *survey sample* group. In contrast, *revealed preference* data on actual purchases are available for both the survey sample and a random sample of the full population. These groups are denoted $RP_{rs}$ and $RP_{ss}$ respectively. Throughout the remainder of the text, we use $D \in \{SP_{ss}, RP_{rs}, RP_{ss}\}$ to index these cases. Note that we *will not* use the selected $RP_{ss}$ data in our main estimation routine. Instead, we link the random sample of actual purchases ($RP_{rs}$) to the survey sample of conjoint choices ($SP_{ss}$), reserving the survey sample of actual choices to assess the degree to which differences between the $RP_{rs}$ and $SP_{ss}$ estimates arise from preference distortions or sample selection.

Let the decision opportunities of consumer $i$ related to dataset $D$ be denoted $t = 1, \ldots, T_i^D$. If one were to estimate preferences based on a single dataset, the underlying utility would be represented as follows:

$$U_{ijt}^D = X_{ijt}^D \, \beta_i^D + p_{ijt}^D \alpha_i^D + \epsilon_{ijt}^D. \tag{4}$$

where the $D$ superscript indicates that the objects may differ across datasets. We now specify why such differences might arise. First, the characteristics themselves ($X_{ijt}$ and $p_{ijt}$) likely differ across the $RP$ and $SP$ contexts. In particular, the $X_{ijt}^{D}$ differ in the set of variables included. The conjoint data include attributes that are common to both the conjoint and actual purchase datasets, along with the hidden and new-to-market attributes, i.e., $K_{Common} + K_{Hidden} + K_{NTMA}$ attributes. The actual purchase data include variables related to the common attributes available in both datasets and variables related to attributes unique to the actual data. This yields a total of $K_{Common} + K_{Actual}$ variables. The $K_{Actual}$ attributes arise because the cardinality of the attribute set for the actual products on offer is often larger than a choice-based conjoint can practically accommodate. Hence, in many conjoint analyses, a set of variables, $\mathcal{K}_{Actual}$, from the vector $X_{ijt}$ are not included in the conjoint data. These missing attributes yield an additional source of randomness that can increase the scale of the errors, $\epsilon_{ijt}^{SP_{ss}}$. Clearly, the preferences for the missing attribute levels cannot be estimated from the conjoint data alone (Islam et al., 2007).

Second, and more controversially, if equation (4) is estimated from conjoint data alone, it will not necessarily match the preferences from equation (1), which are assumed to capture true "purchase preferences", as opposed to what consumers may view or state their "purchase intentions" to be. Relying on the randomly sampled actual purchase data, we assume that the parameters $\alpha_i^{RP_{rs}}$ and $\beta_i^{RP_{rs}}$ represent the (true, revealed) preferences for the attributes and attribute levels observable in the market for the overall population. In contrast, the conjoint estimates, $\alpha_i^{SP_{ss}}$ and $\beta_i^{SP_{ss}}$, represent the stated preferences given the choice tasks, and are observed for the survey sample alone.

Note that these differences between datasets are for the individual-level parameters, while our estimation goal is to recover the *distribution* of preferences across consumers,

$$(\beta_i^D, \alpha_i^D)' = \Delta_0^D + Z_i \Delta^D + \nu_i^D. \tag{5}$$

where $\nu_i^D$ is distributed multivariate normal with variance $\Sigma^D$. For applications like ours, where we aim to vary prices across locations, we require the distribution of the $Z_i$

variables to have meaningful variation across locations. Next we discuss how we link the datasets in order to obtain accurate preference estimates for all parameters.

*Linking Stated Choices And Actual Purchase*

In order to estimate the full set of preferences, we aim to recover these parameter estimates using both the $SP_{ss}$ and $RP_{rs}$ data. Theoretically, one could combine these samples into a single dataset, simply leaving out (setting to 0) the attributes (attribute levels) that are absent from a given observation type and forcing the common covariates to have the same preference parameters. However, in combining stated and revealed preferences, one needs to address three limitations:

1 Choice predictability: Actual choice prediction may exhibit more or less variability (error scale) than hypothetical survey choices. This could arise from the number of tasks, the number of options in the tasks, or the task environment (e.g. the reduced set of attributes in the conjoint tasks). These choice predictability differences can be captured by allowing the scale parameter to differ across the RP and SP contexts (Louviere et al., 2000).

2 Sample selection: Respondents to conjoint surveys are disproportionately drawn from selected types of (heavier) users of the category. They may have more experience making the relevant trade-offs (Ben-Akiva et al., 2015) or have stronger tastes for the products considered. This could cause the conjoint preference distribution to differ from the population distribution in systematic ways, e.g., the unconditional distribution shifts upward. This selection bias can be corrected by conditioning on the observable variables that drive the selection (Feit et al., 2010).

3 Choice context: Actual choice contexts may differ from the artificial context of survey choices (Allenby et al., 2005). These differences can arise directly from the decision context (store location, signage, physical product vs. simple survey design with radio buttons) or indirectly through the choice context inducing different information processing modes or providing other incentives. For example, the hypo-

thetical nature of the conjoint tasks could result in a higher level of construal when making decisions (Trope and Liberman, 2010) that alters the nature of preferences. Likewise, social desirability and related motivations may lead individuals to choose options in a survey that they might not normally consume in practice, such as a choice that serves the public good or a healthier food option (Whitehead et al., 2008). Such choice context effects could cause the conjoint preferences to differ from the true preferences on some attributes or attribute levels. For instance, the outside option might be viewed differently between the conjoint tasks and actual purchases (due to the absence of a clear budget constraint) or price sensitivity may appear different for similar reasons (Horsky and Nelson, 1992).

To effectively integrate the conjoint and purchase data, we need to address these three limitations. First, we discuss the predictability differences, which lead to a common scaling difference between the datasets. The primary strategy for matching scales across datasets involves incorporating a positive scaling parameter $\lambda$ (Louviere et al., 2000) that adjusts the relative variances (only the relative scale is identified). The scaling factors between the $RP_{rs}$ and $SP_{ss}$ datasets are operationalized as

$$[\beta_i^{RP_{rs}} \ \alpha_i^{RP_{rs}}] = \lambda[\beta_i^{SP_{ss}} \ \alpha_i^{SP_{ss}}], \tag{6}$$

which can then be translated into the parameters of the distribution of consumer preferences. This becomes

$$\begin{aligned} \Delta_0^{RP_{rs}} &= \lambda \, \Delta_0^{SP_{ss}} \\ \Delta^{RP_{rs}} &= \lambda \, \Delta^{SP_{ss}} \\ \Sigma^{RP_{rs}} &= \lambda^2 \, \Sigma^{SP_{ss}}. \end{aligned} \tag{7}$$

Second, we turn to sample selection. Our primary means of addressing this problem is to incorporate additional observable variables into the $Z_i$ matrix of consumer types. This is an extension of the demographics-based use of $Z_i$ (Feit et al., 2010). In typical settings, the $Z_i$ may include demographic variables such as income and family size brackets. We extend this approach to include measures of past purchase behaviors such as the Recency, Frequency or Monetary value (RFM) of past category purchases. Since invitation to the survey sample is based on these past purchase behaviors, we can use these variables to construct consistent conditional means (along the primary basis of selection). If we recover the correct relationship at each point along the basis of selection, we can then project these preferences to the population distribution of preferences. Central to this approach is that we need full coverage of the variables that determined who completes the survey sample. Since in most conjoint surveys the invitation probability is likely to be a function of these RFM variables, we expect ex ante that selection biases should diminish sharply by incorporating these variables as observable heterogeneity.[5] Moreover, we can then evaluate this claim using the $RP_{ss}$ data.

Finally, we turn to the context effects. Louviere et al. (2000, Chapter 13) argue, and provide some empirical evidence, that the preferences from different data sources are often equal up to a scaling factor (the $\lambda$ discussed above). However, this approach imposes a number of related assumptions. First, with only a single multiplicative scaling factor $\lambda$, preferences are thereby forced to share the same sign across the $RP$ and $SP$ datasets. Second, using a single multiplicative scaling factor implies that the relative attribute tradeoffs are also common across datasets, which is closely related to the so called "preference regularity" assumption. This implies that 'conjoint preferences' cannot be distorted and instead must match exactly with the underlying purchase preferences. For example, for the median individual in $RP_{rs}$ and $SP_{ss}$ datasets, $Z_{rs} = Z_{ss} = 0$, so that the ratio of brand $b$ to price $= \beta_{i,b}^{RP_{rs}}/\alpha_i^{RP_{rs}} = \lambda\beta_{i,b}^{SP_{ss}}/\lambda\alpha_i^{SP_{ss}}$ must remain the same. While multiple studies have argued in favor of a strong preference regularity assumption (see Louviere et al. 2007 for a review), other studies have rejected the hypothesis that the preferences are equal (Hensher et al. 1998; Swait and Andrews 2003; Brooks and Lusk 2010). Ultimately, this is an empirical question.

[5]The measures of observable heterogeneity are computed based on past purchases before the start of the data. In that sense, current purchases $y_{it}$ are functions of past purchases $y_{i0}$. While the estimated $\Delta^D$ on these $Z_D$ variables should not be interpreted causally, they still provide useful estimates for predicting future demand.

To accommodate potential differences in average preferences revealed across datasets, we introduce additive "mean preference shifters" to equation (3).

(8)
$$\Delta_0^{RP_{rs}} = \lambda \left(\mu_0 + \Delta_0^{SP_{ss}}\right)$$
$$\Delta^{RP_{rs}} = \lambda \left(\mu + \Delta^{SP_{ss}}\right)$$
$$\Sigma^{RP_{rs}} = \lambda^2 \, \Sigma^{SP_{ss}}.$$

The vectors $\mu_0$ and $\mu$ shift the preferences for the "average" individual between the $RP_{rs}$ and $SP_{ss}$ datasets. For instance, an additive mean shifter for the outside good shifts all the inside options in the $RP_{rs}$ dataset, relative to the inside goods in the $SP_{ss}$ data. It also allows relative preferences to differ across datasets. Thus, the mean shifters $\mu_0$ and $\mu$ allow for changes in signs and attribute tradeoffs between the $RP_{rs}$ and $SP_{ss}$ contexts. We further note that these mean shifters can also absorb any sample selection biases that are unobserved (i.e., not captured by the observed heterogeneity variables discussed above).

Thus, to link the $SP_{ss}$ and $RP_{rs}$ data we allow the scale to adjust, incorporate observable heterogeneity in the form of demographic and RFM variables, and allow some subset of preferences to have mean shifts between the two datasets. This last part is most novel here, and we turn now to explaining how the choice information is combined across datasets to accommodate these mean shifters.

*Selection Of The Scaling Parameters And Identification*

We begin by discussing the practical aspects surrounding the set of mean shifters to include in the model. First, notice that we can only apply the mean shifters, $\mu_0$ and $\mu$ to variables in the $\mathcal{K}_{Common}$ set (as these are the only ones for which both datasets are informative). Second, in order to pool information across the two datasets, we must force equality to hold for a non-empty set of the parameters on these $\mathcal{K}_{Common}$ attributes (up to the scaling parameter, $\lambda$). Third, the parameter set that is forced to equality should be (relatively) free of preference distortions. If a distorted preference parameter is instead assumed common, pooling will then infect the other (non-distorted) parameters.

To identify distortion in the conjoint parameters, one can certainly rely on managerial experience, theory, or prior empirical findings. For example, past research and theory suggests alternative specific constants may differ (Louviere et al., 2000), price sensitivity likely differs (Horsky and Nelson, 1992), and that other parameters might differ due to question context or social desirability bias (Whitehead et al., 2008). One can also leverage empirical information from the setting at hand by inspecting the estimates from calibrating the preference parameters separately on the two datasets or by applying a series of likelihood ratio tests for whether to pool on specific parameters or not (Brownstone et al. 2000). We also suggest that, theoretically, one could select the set of attribute variables to mean shift via a LASSO-type routine that uses cross-validation over holdout samples from both the conjoint and actual purchase data (unfortunately, we found this to be computationally impractical in our empirical setting).

Before turning to the formal identification requirements, we offer a brief discussion highlighting the intuition for how the datasets are pooled. The preference parameters that are assumed equal provide the link between the datasets. The pooled preference parameters for the common attributes generate a benchmark magnitude. All other parameters within a dataset then have their magnitudes defined relative to that benchmark. Hence, the relative magnitude of one attribute in the conjoint, say organic, is relative to the common pooled parameter, say for fat level. In this way, by pooling on attribute preferences that are believed to be distortion free, one can obtain the full set of parameters distortion free. In a sense, this essentially amounts to "dummying out" the distorted components. To do so requires including shifters for the parameters with distorted preferences in the conjoint data, so that distortions in the shared attribute preferences do not infect the other preference estimates.

As an example, assume an actual vehicle-purchase dataset contains information about prices, cost per mile, and a number of other attributes unique to the actual purchase dataset and a conjoint purchase set has information about prices, mileage costs, and includes a new electric vehicle. If the conjoint price preferences are believed to be biased and but mileage cost is distortion free, then pooling could occur on the cost per mile attribute. In this case, within the actual purchase data, the calibration of the price sensitivity will be relative to the mileage cost preference *within* the actual purchases. Similarly, for the conjoint data, the evaluation of the new electric vehicle will be relative to cost per mile. Hence, the key relationships to estimating accurate trade-offs are the relative comparison of cost per mile to price and cost per mile to the new electric vehicle. Of course, in practice, more than one attribute may be used for pooling, and the parameters can affect both the intercept and the $Z_i$ coefficients. Although this makes the estimation a more complicated projection, the intuition remains the same.

Turning to identification, given equations (4), (5), and (8), we can identify up to $(N_Z+2)*K_{Common}$ scaling constants (additive or multiplicative) for the common attributes between the $RP_{rs}$ and $SP_{ss}$ datasets. Going back to equation (8), take attribute $k$ that is present in both $RP_{rs}$ and $SP_{ss}$ data. Then, for $k$,

$$\beta_{ik}^{RP_{rs}} = \Delta_{0,k}^{RP_{rs}} + Z_{i,rs}\Delta_{k}^{RP_{rs}} + \nu_{i,k}^{RP_{rs}}, \;\; \nu_{i,k}^{RP_{rs}} \sim \mathbb{N}(0, \Sigma_{k,k}^{RP_{rs}})$$
$$\beta_{ik}^{SP_{ss}} = \Delta_{0,k}^{SP_{ss}} + Z_{i,ss}\Delta_{k}^{SP_{ss}} + \nu_{i,k}^{SP_{ss}}, \;\; \nu_{i,k}^{SP_{ss}} \sim \mathbb{N}(0, \Sigma_{k,k}^{SP_{ss}})$$

We can therefore match the intercepts $\Delta_{0,k}^{RP_{rs}}, \Delta_{0,k}^{SP_{ss}}$, the $N_Z$ parameters $\Delta_{k}^{RP_{rs}}, \Delta_{k}^{SP_{ss}}$ relating observable heterogeneity $Z$ with attribute $k$, and the unobserved heterogeneity variances $\Sigma_{k,k}^{RP_{rs}}, \Sigma_{k,k}^{SP_{ss}}$, giving $N_Z + 2$ separate scaling constants for attribute $k$. To fully saturate the matching requires all $(N_Z + 2) * K_{Common}$ scaling constants, but doing so does not allow pooling between the datasets. To allow pooling, one needs to force at least one parameter to be fixed across datasets.

We note that the more parameters that are restricted to be the same across datasets, the more the information on relative preferences can be used to inform the linkage between the datasets. This implies a tradeoff between (1) including more restrictions (and minimizing the bias from preference distortions that can infect the other estimates) and (2) including fewer restrictions (but increasing the sampling error from obtaining insufficient information to link the datasets effectively). As noted above, in practice, this requires judgment guided by past research, theory, and empirical insights. Again, in priniciple this could be done via LASSO, but with the model and sample sizes considered here we found this approach to be computationally impractical.

These identification requirements are assuming the typical scale and additive constant normalizations that are standard in discrete choice settings. One dataset will have its scale normalized (in our case, the actual data). Second, the additive constant in the actual purchases depends on the definition of the outside good. If the outside good is another choice in the category, then a price index can be constructed for this option and the constant normalized to 0. If the outside option is instead not purchasing at all (as typical in conjoint), then the outside option can simply be normalized to 0, as usual.

One final note is how we use our additional $RP_{ss}$ data (actual choices for the survey sample) to distinguish sample selection biases from contextual biases. The additive and multiplicative scaling constants $\mu_0$, $\mu$, and $\lambda$ aim to avoid distortions that bias the estimates when pooling the $RP_{rs}$ and $SP_{ss}$ data. However, to distinguish between how much role selection bias plays as compared to contextual bias, we use the actual purchase data for the survey sample ($RP_{ss}$) to quantify the role of sample selection due to survey participation. Similar to the formulation of Equation (8) above, we can estimate separate survey sample scaling and shift parameters $\mu_0^{RP_{ss}}$, $\mu^{RP_{ss}}$ and $\lambda^{RP_{ss}}$ between $RP_{ss}$ and $SP_{ss}$, and $\mu_0^{RP_{rs}}$, $\mu^{RP_{rs}}$ and $\lambda^{RP_{rs}}$ between $RP_{rs}$ and $SP_{ss}$ datasets. Since $\mu_0^{RP_{ss}}$ $\mu^{RP_{ss}}$, and $\lambda^{RP_{ss}}$ are estimated between the revealed preference and stated preference data for the *same set of individuals*, they represent the bias between the actual and survey *contexts*, as opposed to *(sub)populations*. Accordingly, differences between, for example, $\mu_0^{RP_{ss}}$ and $\mu_0^{RP_{rs}}$ would then reflect the amount of selection bias (after controlling for the observable RFM heterogeneity) between the random and survey samples.

*Estimation*

Using equations (2), (4), we can write the likelihood of a single choice occasion as

$$P^D_{ijt} = \frac{\exp(X^D_{ijt}\beta^D_i + p^D_{ijt}\alpha^D_i)}{1 + \sum_{j \in C^D_{it}} \exp(X^D_{ijt}\beta^D_i + p^D_{ijt}\alpha^D_i)}, \quad i \in D \tag{9}$$

We note that in our application, we use a single multiplicative scaling constant $\lambda$ and additive shifters $\mu_0$ only for the mean preferences $\Delta_0$ and set additive shifters $\mu = 0$ for demographic/past usage variables. As a result, the largest set of aggregate parameters are $\Theta = \{\Delta_0^{SP_{ss}}, \Delta^{SP_{ss}}, \Sigma^{SP_{ss}}, \lambda^{RP_{ss}}, \mu_0^{RP_{ss}}, \lambda^{RP_{rs}}, \mu_0^{RP_{rs}}\}$. Of course, in most of our models we estimate only a subset of these total parameters. We denote the product purchased by individual $i$ at visit $t$ by $[j]$. The individual data likelihood is in equation (10), the mixing distributions in equations (11) to (14), and the total data likelihood in equation (15), noting that the products for $SP_{ss}$, $RP_{ss}$, and $RP_{rs}$ samples are included only when relevant for the model estimation:

$$\mathcal{L}^D_i(\Theta) = \int_{i \in D} \prod_{t=1}^{T^D_i} P^D_{i[j]t}(\Theta, \nu^D_i) f(\nu^D_i) d\nu^D_i \tag{10}$$

$$(\beta_i^{SP_{ss}},\ \alpha_i^{SP_{ss}})' = \Delta_0^{SP_{ss}} + Z_{i,ss}\Delta^{SP_{ss}} + \Sigma^{SP_{ss}} u_i^{SP_{ss}} \tag{11}$$

$$(\beta_i^{RP_{ss}},\ \alpha_i^{RP_{ss}})' = \lambda^{RP_{ss}}\ (\mu_0^{RP_{ss}} + \Delta_0^{SP_{ss}} + Z_{i,ss}\Delta^{SP_{ss}} + \Sigma^{SP_{ss}} u_i^{RP_{ss}}) \tag{12}$$

$$(\beta_i^{RP_{rs}},\ \alpha_i^{RP_{rs}})' = \lambda\ (\mu_0 + \Delta_0^{SP_{ss}} + Z_{i,rs}\Delta^{SP_{ss}} + \Sigma^{SP_{ss}} u_i^{RP_{rs}}) \tag{13}$$

$$u^D_{ik} \sim \mathbb{N}(0,1) \quad k = 1, \ldots, K \tag{14}$$

$$\mathcal{L}(\Theta) = \prod_{i \in ss} \mathcal{L}_i^{SP_{ss}}(\Theta) \cdot \prod_{i \in ss} \mathcal{L}_i^{RP_{ss}}(\Theta) \cdot \prod_{i \in rs} \mathcal{L}_i^{RP_{rs}}(\Theta) \tag{15}$$

Using $R$ simulated draws of $u_i$ for each individual $i$ from a standard normal (multivariate normal with identity covariance matrix), one can approximate the likelihood defined by equations (10)-(14) as

$$\mathcal{L}^D_i(\Theta) \simeq \frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T^D_i} P^D_{i[j]t}(\Theta, u^{D,r}_i). \tag{16}$$

We estimate the above system using Simulated Maximum Likelihood Estimation (SMLE) with $R = 5000$ random draws.

## *DATA AND SETTING*

Our application is to single-cup Greek yogurt in a leading grocery retailer observed from February 6 to October 15 2011. Greek yogurt sales grew from 3% of all yogurt unit sales in 2008 to 18% in 2012. The growth leading up to 2011 was largely believed to be fueled by the entry of Chobani, the leading national brand. The successful growth of Chobani led the retailer to launch its own private label Greek yogurt product. We study the period after Chobani's entry (January 2011) to after the retailer launched their private brand in this category, which was roughly nine months following the Chobani launch.

The retailer had a long-running and successful private label program in place in other categories. At the planning stage for its private label launch in Greek yogurt, the retailer was debating what line of products to offer in the category, including a potential "high-end" offering that would include additional nutrients or be organic. Preferences for these attributes, along with the private brand itself, were not yet observable in the market place. To support managerial decisions related to product selection and the launch (e.g., pricing), the retailer undertook a conjoint analysis that involved sending online surveys to members of their loyalty card club in May 2011. Below we describe the loyalty card data, as well as the related conjoint design and dataset.

### *Loyalty Card Data*

The loyalty card data consists of actual purchases/revealed preferences for a random sample of 4,288 individuals and survey sample of 510 individuals ($RP_{rs}$, $RP_{ss}$ and $SP_{ss}$ datasets respectively) between February and October 2011 across 79 stores of the supermarket chain. Figure 1 shows the relationship between the samples and datasets.

The period between February and October 2011 was marked by expansion in Greek yogurt share (of total yogurt sales), and multiple brand entries and exits. Chobani and Dannon entered the retailer's stores in January 2011, just prior to our analysis period.

Brown Cow exited Greek yogurt in July 2011, and the private label Greek yogurt was introduced in September 2011. Figure 2 shows these changes in the Greek yogurt market (across the retailer's stores) for the random and survey sample individuals.

Figure 2 shows the time series of Greek yogurt purchase shares for the $RP_{rs}$ and $RP_{ss}$ datasets. This share is measured as the share of Greek yogurt in total single cup yogurt unit sales; throughout, we treat club-pack purchases as multiple single-cup purchases. The $RP_{rs}$ data Greek yogurt market shares increased from 16% in February to 28% by July, 2011, whereas the survey sample shares rose from 47% in February to 69% in the same period. These differences reflect the fact that the survey was sent to a selected group of individuals that were more likely to have bought Greek yogurt in the last three months than the overall population of chain customers. Figure 2 also depicts the timing of our estimation and forecast periods. We split the total February-October 2011 period into three groups: estimation period (24 weeks between February 6-July 23, 2011) before Brown Cow exit, forecast period 1 (6 weeks between July 24-Sept 3, 2011) after Brown Cow exit and before private brand entry, and forecast period 2 (6 weeks between Sept 4-Oct 15, 2011) after the private label entry. This will allow us to evaluate the performance of the model in a setting with no new-to-market attributes (period 1) as well as one that includes them (period 2).

Recall that the probability that a loyalty card holder was invited to the survey depended on their past purchase of Greek yogurt. Our approach for correcting for this selection involves projecting the tastes for these heavy users to those with fewer or no recent purchases of Greek yogurt (i.e., to the full population, as represented by the random sample of loyalty card holders). This clearly requires having some individuals in the survey sample that span the range of potential RFM variables. Figure 3 reveals that we indeed have some individuals throughout the distribution but, as expected, our sample is clearly slanted toward higher purchase recency, frequency, and monetary values. We note that this is typical in conjoint surveys where participation is selected on past purchases (Ben-Akiva et al., 2015).

Table 1 compares the revealed preference brand shares (of sold units) for the random sample ($RP_{rs}$) and survey sample ($RP_{ss}$) in the estimation period. During these 24 weeks, we can see that survey sample individuals are three times more likely to purchase Greek yogurt than random sample individuals (67.28% vs 20.14%). However, conditional on purchase, the survey sample individuals have similar brand shares to the random sample (cannot reject the null in a Chi-Square Goodness-of-Fit Test; test-statistic = 1.50).

The random sample and survey sample individuals further differ in their demographic and usage variables. We use three measures of past category usage - days since last Greek yogurt purchase (Recency), number of store visits with Greek yogurt purchase in past 90 days (Frequency), average expenditure in Greek yogurt per store visit in past 90 days (Monetary value). These RFM measures are initialized at the start of the data period, and therefore are time-invariant during the analysis. We also include two pure demographic variables - whether the household annual income is below $\$75,000$, and whether the consumer belongs to a small family (with less than 3 members). Table 2 shows the statistics for the RFM and demographic variables for the $RP_{rs}$ and $RP_{ss}$ data at the start of February, 2011. By the time the conjoint survey was launched in May 2011, the RFM variables for the survey sample individuals had changed further, as is shown in the last column. To facilitate interpretation of the $\Delta_0$ parameters as reflecting the "average" person in each dataset, the observable heterogeneity matrices $Z_{ss}$ and $Z_{rs}$ are mean-centered and do not include an intercept.

To ensure similar weighting of information across $RP_{rs}$, $RP_{ss}$ and $SP_{ss}$ datasets, we select 510 individuals from the random sample for estimation purposes. These individuals were selected to have at least one store visit in the estimation and two forecast periods. Table 2 reveals that this mild selection poses little threat to the representativeness of the sub-sample of individuals. In a similar vein, we use revealed preference $RP_{ss}$ data for the survey sample individuals before the release of the conjoint survey (13 weeks from Feb 6-May 19, 2011), to guard against the possibility that preferences change systematically after answering the survey (refer to Figure 2). Since survey sample individuals are, on average, more frequent patrons of the supermarket, using only 13 weeks for the $RP_{ss}$ data produces an average number of choice occasions per individual that is similar between the $RP_{rs}$ (56.34) and $RP_{ss}$ data (58.81).

The $RP_{rs}$ and $RP_{ss}$ estimation samples have $28,731$ and $29,640$ purchase occasions respectively. Here, we define a 'purchase occasion' as an instance where the consumer chooses a product $j$. During a single shopping trip, consumers choose between an outside option and up to 14 available single-cup Greek yogurt options. We identify the available products based on non-zero store sales from aggregate data and treat multiple single-cup or club-pack purchases within the same shopping trip as a series of independent single-cup choices. The choice options vary on dimensions such as brand, whether the yogurt is zero-fat or low-fat, is plain or fruit-flavored, whether the brand was purchased in the previous shopping trip (state dependence), and the price (calculated as the average store-week price from aggregate data). Following Dubé et al. (2010), state dependence is calculated only for the inside options, as an indicator of the last brand purchased that increments that brand's utility. We also include state dependence in the conjoint data, which is based on the respondent's last purchase prior to the survey, similar to Swait and Andrews (2003). Table 3 contains descriptive statistics for the $RP_{rs}$ and $RP_{ss}$ choice sets.

*Conjoint Data*

The conjoint survey was launched in May 2011, with over 2200 invited respondents (refer to Figure 2). The customers invited to participate in the online survey were chosen to have at least one purchase in Greek yogurt in the past 6 months. Of these invited participants, 510 respondents completed the full survey within a week of its launch. Participants were offered the chance to win a $100 gift card to the retailer for their participation in the survey.

The conjoint survey was designed as a choice-based conjoint questionnaire fielded via Qualtrics. Respondents were initially asked their propensity to consume Greek yogurt, what attracts them to the category, and how they use it (cooking/snacking). Following these questions were some generic instructions about the conjoint tasks including, "Now you will be asked to select between multiple options of Greek yogurt. For each of these choices, assume your preferred flavor(s) are available within each of the options presented." These instructions were intended to avoid including the large variety of flavor options. Seven different sets of 12 tasks were designed using Sawtooth Software's design module and were assigned randomly to respondents. Each task had four profiles of Greek yogurt and a "None-Of-These" option. Appendix figure 6 shows an example question.

The profiles varied attributes such as brand, fat content, plain/fruit flavors, organic/all natural content, healthy attributes (probiotic/vitamins/omega-3/fiber), packaging styles and price. The full set of attributes is listed in appendix table 14 along with the set of attributes in the marketplace. We characterize these based on the hidden, new-to-market, actual, and common designations defined in section 2. In the actual marketplace, consumers could choose over brand, fat content, fruit/plain, cup styles, and price. Some brands only carried organic or all natural options. For example, Oikos was the only organic brand and it carried only organic products. Hence, the preference for these attributes was confounded with brand preferences in actual choices, i.e., the hidden attributes. The new-to-market attributes (NTMA) include healthy ingredients (probiotic/vitamins/omega-3/fiber) as well as a new attribute level for the private label brand and new packaging styles. The conjoint questionnaire, therefore, elicits preferences for these hidden and NTMA attributes which cannot be obtained from revealed preference data. The only attribute level in the actual data that was not included in the conjoint analysis was the Brown Cow brand value.

## *RESULTS*

We divide this section into four parts. First, we present baseline results for choice models estimated separately, that is using the actual purchase ($RP_{rs}$) and conjoint ($SP_{ss}$) data in isolation. This initial exercise provides insight into the overall sample selection and contextual bias effects that must be accounted for in our proposed data-fusion model. Second, we discuss the results of our preferred model (henceforth, $CPC$, for 'consumer panel conjoint') and compare it to a strong benchmark (henceforth, $BENCH$) that incorporates the innovations introduced in both Swait and Andrews (2003) and Feit et al. (2010), along with the standard scale adjustment parameter. In particular, the $BENCH$ model includes both state dependence (as suggested by Swait and Andrews (2003)) and individual-level demographics serving as observable heterogeneity (as used by Feit et al. (2010) to address sample selection bias), in addition to a constant to scale the error variance between the $SP_{ss}$ and $RP_{rs}$ contexts (to account for predictability differences). Third, we compare each model's predictive validity across the estimation and holdout datasets, and over the three different time periods (recall Figure 2). Finally, we explore in more detail the role of selection and contextual biases and how our approach addresses each. This last exercise leverages all three datasets, the random sample ($RP_{rs}$) and survey sample ($RP_{ss}$) purchase data, as well as the survey sample conjoint data ($SP_{ss}$) to clarify the magnitude of the two forms of bias. We also examine the empirical impact of including the demographic and usage variables.

*Separate Sample Estimation*

To begin, we estimate using the $RP_{rs}$ and $SP_{ss}$ data separately following equations (4), (5), (10), and (14). The results are presented in Table 4. We present estimates corresponding to the preferences of the average individual, $\Delta_0$, and the standard deviations for the unobserved heterogeneity, $\sigma$. The coefficients, $\Delta$, are presented in appendix tables 15 and 16. Rather than interpreting coefficients one at a time, we focus on identifying potential distortions in the $\Delta_0$ parameters as a whole.

To analyze potential preference distortions, Figure 4 provides a scatter plot of the mean parameters for the common attributes, comparing the $SP_{ss}$ and $RP_{rs}$ models. If the parameters represent the same relative trade-offs, then points far from a line passing through zero represent preference distortions (Swait and Louviere, 1993). As a visual aid, we include a scaling line based on the optimal value for a single scaling $\lambda$ (see the $BENCH$ model estimates below). Apparent from the chart are three attribute parameters that are far from this line–the coefficient on a dummy applied to all inside goods, the price sensitivity parameter, and the state dependence coefficient. We now interpret these preference distortions and one additional distortion in the conjoint estimates.

First, note that the inside options in the $RP_{rs}$ data have a large negative intercept (mean = -8.68, se = 0.19) whereas the $SP_{ss}$ data shows the inside options are preferred to the "None-of-These" option (mean = 2.11, se = 0.25). These differences could arise from either contextual or selection biases, or both. For instance, one possible issue is apparent from appendix figure 6. Consumers may interpret the "None-of-These" options differently in the survey (i.e., to be no purchase), whereas they would instead choose a yogurt other than Greek in the actual context. Hence, they might respond with a lower probability to "None-of-These" because of the question format (or the lack of a clear role for a budget constraint). Alternatively, these differences may arise because the survey sample consumers are a selected subset of the full population who have stronger preferences for Greek yogurt. In section 4.4, we explore which of these explanations is most likely, but for now we abstract from this finer distinction and simply note the distortion.

The next two major differences between the two sets of estimates include the parameters on price and state dependence, which reflect magnitude differences. First, price sensitivity is significantly lower in the conjoint data ($RP_{rs}$: mean = -10.08, se = 0.20; $SP_{ss}$: mean = -2.64, se = 0.20). This is consistent with the results of Horsky & Nelson (1992), who found that consumers attend more carefully to price when faced with actual choices rather than conjoint tasks. Second, state dependence plays a bigger role in the conjoint survey (mean = 2.52, se = 0.15) than in the $RP_{rs}$ data (mean = 0.84, se = 0.06), especially relative to price. Since the state dependence dummy is triggered only for the inside options, the lower state dependence parameter indicates higher substitution with the outside good in actual purchases as compared to the survey choices.

Finally, we identify one other distortion related to the fruit/flavor attribute, which is not immediately apparent in figure 4. Recall, the conjoint survey asked consumers to assume their "favorite flavor was available" while evaluating their preference for plain vs. fruit options. This aggregation in the question is potentially problematic as it might artificially reduce the preference for flavor variety. To evaluate this possibility, we examine the relative trade-offs of fruit with other attributes. We illustrate with the comparison to the zero-fat attribute. While fruit options are more than 2.5 times preferred compared to zero-fat options in the $RP_{rs}$ data, the preference ordering is reversed in the conjoint survey. Thus, as one might expect from the question's forced aggregation, the fruit preference is greatly understated in the conjoint preferences. Such a drastic relative preference reversal could clearly lead to distortion problems.

These documented distortions inform our modeling choices regarding which attributes to include in the mean-shift adjustment set. Based on the above analysis, we allow for additive shifters to the collection of inside goods, price sensitivity, state dependence and fruit flavor parameters in our proposed model. These additive shifters correspond to the attributes revealed to have the most critical distortions. As noted in the model section, the modeling choice for how many and which attributes to give additive shifters balances the desire to correct problematic differences between the contexts (or samples) with the need to provide a basis upon which to pool the datasets in order to forecast attributes not yet observable in the marketplace. We take a relatively parsimonious stance in our application, yet as we will show below, even just these four additive shifters allows us to better capture the trade-offs in both datasets and improve forecasts both within and outside the estimation sample.

Finally, in both estimations, the inside options have an intercept, and all the brand estimates represent deviations from it, with Dannon being the excluded brand. However, it should be noted that the brand intercepts in the $RP_{rs}$ and $SP_{ss}$ standalone estimations should be interpreted differently. In the $RP_{rs}$ standalone estimation, attributes such as organic/all natural/rBST-free are prefectly correlated with the brand intercepts. Hence, the brand intercepts in this estimation also subsume the preferences for these attributes. These hidden attributes are incorporated in the $SP_{ss}$ parameters in figure 4.

*Revealed Preference-Stated Preference Matching*

We now consider our preferred model, which integrates both the revealed preference and stated preference data across the $RP_{rs}$ and $SP_{ss}$ samples. Our proposed model ($CPC$) from equation (8) uses a single multiplicative scaling constant $\lambda$ and additive mean shifters for the inside good, fruit flavor, price sensitivity, and state dependence. We compare our model with the strong benchmark model ($BENCH$) that is nested in our model. The $BENCH$ model has a single scaling parameter, but no additive shifters. As before, we report in table 5 the mean preferences $\Delta_0$ and unobserved heterogeneity standard deviations $\sigma$, along with the scaling constants $\lambda$ and $\mu_0$. Detailed results are available in appendix tables 17 and 18.[6]

On comparison, $CPC$ fits the data much better than $BENCH$ in terms of total likelihood (likelihood-ratio test rejected at $p < 0.01$), as well as Akaike and Bayesian information criteria. Further, we note that $CPC$ performs better in both the $SP_{ss}$ and the $RP_{rs}$ data. This improved performance stems from the greater flexibility our model affords.

Since the parameters reported in table 5 are in $SP_{ss}$ scale, they can be directly compared with the standalone $SP_{ss}$ results in table 4. On the whole, the $CPC$ estimates are closer to the standalone $SP_{ss}$ estimates than are the $BENCH$ estimates. For instance, the $CPC$ estimate for the inside good is 2.64 (se = 0.22), while the corresponding $BENCH$ estimate is -1.81 (se = 0.09), relative to the $SP_{ss}$ estimate of 2.11 (se = 0.22). This closer relationship arises because the additive shifters allow more degrees of freedom in fitting the $SP_{ss}$ data, rather than restricting these parameters to match across the two datasets.

The more important comparison, however, is on the $RP_{rs}$-scale, as this is the scale used to predict actual (future) choices. Converting to the $RP_{rs}$-scale, the $CPC$ parameters again perform better than those of the $BENCH$ model. For instance, the estimate for the mean inside good preference in the $RP_{rs}$ scale for $CPC$ is $(2.64-7.20)*1.91 = -8.70$, much closer to the standalone $RP_{rs}$ estimate of $-8.68$. In comparison, the corresponding $BENCH$ estimate is $-1.81 * 1.80 = -3.26$. Similarly, the $RP_{rs}$ price-sensitivity (mean = -10.08, se = 0.20) is closer to the $CPC$ estimate (mean = -7.44, se = 0.23) than the $BENCH$ estimate (mean = -5.36, se = 0.19). Figure 5 depicts scatterplots of the full set

[6]We also estimated models where $Z_i$ included only demographics variables. The reported models with the RFM variables significantly improve over the demographics only versions, so we focus on the reported models. In section 4.4, we discuss the benefits of adding the RFM variables.

of common parameters for the $BENCH$ and $CPC$ models, respectively. As apparent, in general, the $CPC$ model parameters more closely match the horizontal line, although this is driven largely by the price sensitivity and inside good parameters. Thus, one can see that our model generally matches more closely the preferences as revealed by the $RP_{rs}$ data than the benchmark model, but that our model's matching is also not perfect. The imperfect match to the $RP_{rs}$ data can be thought of as a cost to integrating the conjoint data and incorporating the NTMA and hidden attributes into predictions, at the expense of sacrificing some fit for the fully observed actual choice data. As mentioned, we take a relatively parsimonious stance on the number of shifters to incorporate, but conceptually, one can include many more to improve the match on common attributes. Such increased flexibility, however, will weaken the foundation linking the datasets.

*Predictive Validity*

In this section, we show that mapping the tradeoffs more accurately also translates into better predictive power. We include validation on the estimation ($RP_{rs}$) sample as well as the remaining random sample individuals (which form the holdout sample) for three different periods depicted in Figure 2 –the estimation period (February 6-July 23, 2011), forecast period 1 (July 24 - Sept 3, 2011) and forecast period 2 (Sept 4 - Oct 15, 2011).

Table 6 shows the predictive log-likelihoods, calculated using $R = 10000$ draws (along with restatement of the estimation log-likelihoods on the $RP_{rs}$). Note that the standalone $SP_{ss}$ model cannot predict in the estimation period, since the conjoint survey did not have an attribute level for the Brown Cow brand. Similarly, the standalone $RP_{rs}$ model cannot predict for forecast period 2, as there is no estimate for the private label brand preference.

We first focus on the periods when the $CPC$ can predict on the $RP_{rs}$ data and compare its performance to the standalone $RP_{rs}$ model, as this is the (gold) standard choice model for recovering revealed preferences (i.e. actual choices used to predict actual choices). By construction, in the estimation period and considering within sample fit, the $RP_{rs}$ data must perform best (as it is not constrained to fit the conjoint data). Notably, the $CPC$

model performs similarly to the $RP_{rs}$ model within sample and performs better in the hold-out sample, whereas the $BENCH$ model performs considerably worse.

After Brown Cow exits (forecast period 1), the $CPC$ model again predicts similarly well to the $RP_{rs}$ model and predicts best in the hold-out sample. Again, the $CPC$ and $RP_{rs}$ models far outperform the $BENCH$ model. But in the period with Brown Cow absent, when the $SP_{ss}$ model can predict actual choices, we can also see that, as one might expect, the conjoint alone predicts worse than either the $BENCH$ or $CPC$ model, due to its preference distortion and selection bias issues.

We now consider forecast period 2, when the private label brand has entered. This is the critical forecast period for evaluating the model's ability to inform managerial decisions. In this period, the $CPC$ model has the best predictive likelihood both within sample and in the holdout, outperforming both the choice-based conjoint model and the $BENCH$ model. Overall, the predictive likelihoods suggest that the $CPC$ model performs similar to the $RP_{rs}$-only model when that model can predict and performs the best outside of that setting.

Table 7 presents the root mean squared error (RMSE) calculated as the difference between the percentage of individual product choices in the sample period versus the forecasted shares for the same period. In this table the aggregation of shares is at the individual level, with the average taken over individuals. Similar to the predictive likelihoods, the conjoint model performs worst, and here we can see that that leads to approximately twice the error as the other models. Thus, incorporating the actual choice data leads to meaningful improvements in prediction. This finding is consistent with past findings regarding data fusion (e.g., Swait and Andrews 2003). For the the estimation period and forecast period 1, the error rates of the $CPC$ model are comparable to those of the $RP_{rs}$ model. Thus again, the $CPC$ model appears able to reproduce the actual trade-offs with a similar accuracy to established revealed preference methods. In contrast, the $BENCH$ model has more than 40% more error, suggesting our model performs demonstrably better. Consistent with the difficulty of more distant forecasts, the error rates rise slightly in forecast period 1 as compared to the estimation period, and then again in forecast period

2. However, $CPC$ still performs better than both the conjoint and the $BENCH$ models, with $CPC$ achieving an RMSE of around 0.45. For RMSE on individual purchase shares, the $CPC$ model performs best on the critical prediction tasks and can reproduce closely the $RP_{rs}$ model on forecasts where new to market attributes are not required.

Finally, we consider the predictions needed for the store-level pricing exercise we will conduct in section 5, i.e., store-level share predictions. These prediction RMSE are calculated analogously to those in Table 7, but where the aggregation is at the store level (instead of individual) and then averaged across stores. Table 8 presents the results. The general pattern follows closely that of the individual-level RMSE results (as one should expect), but here the differences are more stark. The error rates for the $RP_{rs}$ and $CPC$ models decrease sharply as errors average over individuals, but the errors in the $BENCH$ and conjoint-only models do not. One can interpret these decreasing errors rates as arising from non-systematic errors in the $CPC$ and $RP_{rs}$ predictions, whereas in the $BENCH$ and conjoint-only models the error rates do not decrease as sharply suggesting more bias than non-systematic error. The results for the hold-out sample in forecast period 2, on which the managerial decisions rely most heavily, are particularly striking: the $CPC$ model generates less than half the level of error of the strong benchmark model, $BENCH$, and less than one fourth the error of the conjoint-only model. This is strong evidence for the predictive benefits of the $CPC$ model, especially in supporting managerial decisions related to store-level pricing.

*Exploring The Roles Of Selection And Contextual Biases*

A key strength of our empirical setting is that we observe actual purchases for both the survey sample and the random sample of customers. This information allows a unique opportunity to quantify the extent of contextual versus selection biases and the degree to which our RFM solution corrects for selection. Because the typical design of conjoint involves inviting or selecting on past purchasers (Ben-Akiva et al. (2015)), and because such selection can lead to biased estimates, this analysis provides a clear window into the effectiveness of using observable covariates to control for survey bias.

The significance of the additive and multiplicative scaling constants in our estimation and forecasting already suggests that there are some distortions in the preferences obtained from the conjoint survey as compared to those from the actual purchase information. We now explore whether these distortions arise from the differences in the two choice contexts (contextual biases), or due to the differences in the consumer types that comprise the $RP_{rs}$ and $SP_{ss}$ datasets (selection biases), and to what extent any selection biases are corrected by the inclusion of the RFM variables as compared with including only the demographics variables. For this investigation, we leverage the information on the actual purchases of the survey sample individuals, i.e, the $RP_{ss}$ data. Since the $RP_{ss}$ and $SP_{ss}$ data are the actual and conjoint responses, respectively, for the *exact same* individuals, preference distortions in this data reflect contextual bias alone, as there is no selection in this comparison (though this sample is a selected subset of the full population). Thus, the additive shifters for distortions between $SP_{ss}$-$RP_{ss}$ reflect pure contextual bias in this subpopulation. As such, by comparing the $SP_{ss}$-$RP_{ss}$ shifters to those for $SP_{ss}$-$RP_{rs}$, we obtain an estimate of the size of the selection bias compared to the contextual bias. If there is no selection bias on unobservables, there should be no difference between the $SP_{ss}$-$RP_{rs}$ and $SP_{ss}$-$RP_{ss}$ shifters.

The focal results of this combined analysis are shown in table 9. Note that the $RP_{ss}$ results (and $SP_{ss}$-$RP_{rs}$ comparison) involves a new estimation, as we have not yet used the $RP_{ss}$ sample in estimation. To assess the impact of different demographic controls, we present results from models where $Z_i$ contains the demographics-only variables and where $Z_i$ contains the full set of variables including the RFM-related ones. We first discuss the results for the model with the $Z_i$ containing the demographics and RFM variables, and then discuss differences between that model and the demographics-only case. Detailed results are included in appendix tables 19 and 20.

To start, the significance of the $\lambda^{RP_{ss}}, \mu_0^{RP_{ss}}$ estimates clearly points to the presence of contextual bias for the same set of survey respondents. We see contextual bias in each of the attributes that we include as shifters. As noted earlier, the largest distortions relate to the inside good (-5.59), price sensitivity (-1.05), and state dependence (-1.41) with the fruit distortion being much smaller (0.64). This is strong evidence that the preference distortions documented earlier are not simply about who is selected into the survey sample.

Moreover, for the Dem & RFM model, the $SP_{ss}$-$RP_{ss}$ and $SP_{ss}$-$RP_{rs}$ shifters are significantly different from each other (each p-value<.01). The additive shifters for the inside good $\mu_0^{RP_{ss}}$ (mean = -5.59, se = 0.27) and $\mu_0^{RP_{rs}}$ (mean = -6.77, se = 0.29) have the largest difference (-1.18), followed by that for state dependence (-0.43), price (-0.25), and finally fruit (-0.14). This indicates that there is significant selection on unobservables even after accounting for the RFM variables. However, in each case, the size of the selection effect is relatively small compared to the total bias. For the inside good, the selection-related part is -1.18 out of a total -6.77 or 17% of the effect size. The effect sizes for the others are all less than 25%. Hence, after controlling for the RFM variables, a relatively low level of selection bias persists.[7]

The model results we have discussed thus far include in the $Z_i$ the RFM and demographics variables. Comparing the shifters for this full set of $Z_i$'s to the shifters for a model with only the demographics can inform the kind of selection that the RFM variables are able to capture. The second set of results in the table present the demographics-only model and the last column compares the two models. We find that the RFM variables sharply reduce the selection effect related to price sensitivity. We find a difference of 1.25 between the $SP_{ss}$-$RP_{ss}$ and the $SP_{ss}$-$RP_{rs}$ shifters in the demos-only model, but only 0.25 difference in the RFM + demos model (p-value<0.01). Thus, consistent with intuition–i.e., that those in the survey sample are more likely to buy Greek yogurt, a relatively expensive product, and as a result are likely to be less prices sensitive–the RFM variables correct for a large part of the selection bias in price sensitivity. In this case, a full 4/5 of the selection bias is captured by the RFM variables, but some remains. Further, in total almost 3/5 of the price sensitivity bias is due to selection. The inside good differences of 1.60 for the demos-only and 1.18 for the RFM + demos model is also directionally

[7] Because state dependence is captured in $X_{ijt}$ (as suggested by Swait and Andrews (2003)), the base utility differs across consumers based on their past purchases. Since past purchases are related to the selection process, including these state dependence variables may already adjust for some degree of selection.

consistent with the expected selection bias (p-value <0.001), but the effect size is much smaller. In fact, the inside good difference, along with both the price sensitivity and state dependence differences, are small enough that the individual point estimates across the two models (e.g., $RP_{ss}$-$SP_{ss}$ for Dem & RFM vs. $RP_{ss}$-$SP_{ss}$ for Dem-Only) are not significantly different from one another (inside good difference = -5.59+5.01 = -0.58, sd = 0.36). This suggests that the main benefits from the RFM variables in terms of handling the selection biases arise from correcting for price sensitivity.

To summarize, after controlling for the RFM variables, the bulk of the remaining bias is context related. In particular, less than 1/4 of the remaining bias appears to arise from selection. However, without controlling for the RFM variables, the selection bias can be as large as 3/5 in the case of price sensitivity. This suggests that simply incorporating the RFM variables in the $Z_i$ is not enough to obtain accurate preference estimates, and that both the inclusion of the RFM variables and the inclusion of the shifters are important to address the biases in the data.

## ***MODEL-BASED MANAGERIAL DECISIONS***

We now present an illustrative application of the $CPC$ model to managerial decisions aimed at optimizing a new product line's bundles of attributes and pricing across locations. For the purposes of this illustration, we focus on decisions to optimize profits for the private label product only (extending this analysis to more products is conceptually straightforward, but requires knowledge of or more assumptions related to the costs). We assume that the retailer plans to launch a line of private label yogurt products that will include both fruit (e.g., blueberry) and non-fruit (e.g., plain) options, and that the marginal cost is \$0.40 for a basic product without organic or other added ingredients (communications with the retailer suggested this would be a reasonable approximation). In what follows, we describe the forecasting process, objective function, and data setting, and in separate sections discuss the product selection and store-level pricing problems.

To forecast demand, we use maximum a-posteriori (MAP) estimates at the individual-level to capture an individual's tastes as revealed by their estimation period purchases.[8] Using the predicted probabilities from the MAP estimates (see appendix 6 for details), we then simulate total demand for potential purchase occasions, as given by

$$Q_j^{MAP}(\Theta, X, p, Z, \{\nu_i^{MAP}\}) = \sum_{i=1}^{N} \sum_{t=1}^{T_i} P_{ijt}(\Theta, X, Z, p, \nu_i^{MAP}), \tag{17}$$

where total predicted demand for product $j$, $Q_j^{MAP}$, is a function of the parameters $\Theta$, attributes $X$, prices $p$, demographics $Z$ and MAP estimates $\nu_i^{MAP}$. We assume a marginal cost $c_{PL}$ for the private label and compute the expected profits from the sale of private label products as

$$\Pi_{PL}^{MAP}(\Theta, X, p, Z, \{\nu_i^{MAP}\}) = (p_{PL} - c_{PL}) Q_{PL}(\Theta, X, p, Z, \{\nu_i^{MAP}\}). \tag{18}$$

In practice, retailers often place constraints on the set of prices considered. Based on communications with our focal retailer, three dominant constraints are of interest. First, brands should be line priced so that all flavors and fruit varieties are priced the same, which we always enforce. Second, the retailer rarely allows the private label to be priced more than \$0.10 below than the comparable national brand (in the same store).[9] This constraint is also consistent with findings in the literature (Geyskens et al. 2010; Ailawadi et al. 2008; Chintagunta et al. 2002), and we require our prices to satisfy this constraint. The third constraint involves zone pricing. This retailer, like many others, sets prices in broad geographic zones. Some of our counterfactual exercises are geared toward exploring this constraint.

As data, we consider the 1495 individuals from the random sample who visited at least one of the chain's 76 stores following the introduction of the private label brand (forecast period 2, see Figure 2). Following Meza and Sudhir (2010), we dropped the first four weeks after private label introduction to ensure that product roll-out, stock-

[8]Another approach to predicting demand would be to directly integrate across all $K$ attribute dimensions. The results from both methods are qualitatively similar.

[9]For the actual private label launch, which was aimed at matching Chobani's line, only 4 stores have prices above this constraint, the median price difference from Chobani is \$0.11 and the minimum other than those 4 stores is \$0.10.

outs, and limited initial product awareness do not adversely affect demand. We use the estimation period purchases to obtain MAP estimates for these individuals. As noted above, we assume that the retailer does not change availability or prices of the other products/brands in its stores. Therefore, the only change to the individual's choice set is the type of private label products they would face (for product decisions), and their prices, $p_{PL}$. Additionally, in calculating shares, we only consider the first store visit for these individuals, ensuring that their level of state dependence (of previous brand purchased) is already known. All profit calculations are for this first visit by this set of customers. Since the same store visit may entail multiple choice occasions, we include 7702 total choices in our counterfactuals.

*Product-Line Decisions*

The manager first needs to decide which attributes to introduce in the new product line for the private label brand. To address this problem, we compare products adjusting one attribute at a time from a baseline product (low fat, rBST-free, with no additional probiotics/vitamins/omega-3/fiber content and normal single-cup packaging). In this first analysis, we do not adjust the costs based on adding potentially more expensive attributes (e.g., additional ingredients), so for these cases the profits are overstated. In the three scenarios regarding packaging styles, we match Fage's policy, so that the plain flavor has the normal cup style and the fruit flavors have the special packaging, which is meant to contain the extra ingredients. Table 10 shows the maximum expected profits and corresponding prices at cost level $c_{PL}$ = \$0.40 for unconstrained prices as well as prices constrained to be at least 10 cents below Chobani.

Table 10 reveals that the expected profit for the baseline private label product at $c_{PL}$ = \$0.40 is \$550.55 at an optimal price of \$1.07. Adding the zero-fat attribute to the product, improves the profits to \$794.17, the highest increment of any single attribute shift. This profit is achieved at a per cup price of \$1.14, implying that a \$0.07 price mark-up is acceptable for introducing the zero-fat attribute. Making the product-line all-natural (expected profits of \$592.67) is preferable to organic (\$581.68) or rBST-free (the baseline, \$550.55). Similarly, adding probiotics, omega-3 and fiber is preferable. Finally, changing the packaging style reduces the expected profits from the baseline. From Table 10, we conclude that the retailer should consider introducing a zero-fat, all-natural Greek yogurt, potentially with additional ingredients, depending on their costs.

We also illustrate a comparison of the suggested basic product-line against an alternative, high-end product that is organic and contains probiotic, omega3, and fiber (as suggested by the output in Table 10). Because the costs of the additional ingredients are not known, we estimate profits for the composite product-line under a range of marginal costs between \$0.45 – \$0.75. For this comparison, we constrain the prices to be within \$0.10 of the highest price brand, noting that with the added ingredients this product might be more attractive than any existing product. Table 11 shows that whether the basic or composite product is preferred depends on the costs. If producing the composite line costs at least \$0.30 more than the basic line (marginal cost of \$0.70), the corresponding expected profit is \$728.13, so that the basic product line dominates. However, if the difference in marginal costs is \$0.25 or less, the composite line dominates.

*Pricing Decisions*

In the previous product-line counterfactual exercise, we used uniform pricing across stores. We now study the benefit from allowing prices to vary by store location. For this analysis, we assume that the retailer introduces the basic product-line with a cost of \$0.40 per unit.

Although many retailers would like to exploit cross-store differences in demand conditions, they are often concerned with consumer and PR backlash over targeted pricing. If consumers shop in multiple stores within the same chain and observe significant price differences across stores, this could negatively impact the chain's reputation and threaten profits. Retailers may seek to mitigate this risk by limiting the scope for cross-shopping consumers to be met with different prices for the same product (across the chain's own outlets). In particular, stores that are geographically close, with more consumers having cross-shopping trips, should be constrained to have similar prices, whereas more isolated stores should have the flexibility to set different prices. In our context stores are sufficiently close to raise concerns. As table 12 shows, around 12.9% of the chain's customers visit more than one it's outlets during the 6 weeks after the private label launch, indicating that cross-shopping is meaningfully large.

To illustrate how cross-shopping can impact pricing and profitability, we impose the following constraints: For each store, less than 10% of its cross-shopping customers should face a price difference of more than 5 cents and less than 5% of these customers should face a price difference of more than 20 cents (for the private label product). These constraints impose a regression-to-the-mean effect, where the highest-priced store within a region (or cluster of cross-shopped stores) might need to price closer to the other stores, or the lowest-priced store might need to price higher.

To implement the cross-shopping constraints, we use the actual stores visited by our random sample of consumers in forecast period 2 to obtain the extent of cross-shopping, and, as a result, the distribution of price differentials across customers. Since these cross-shopping estimates are constructed from a random sample, this measure of cross-shopping is also representative of the actual correlation of shopping trips across stores.

The results for the various pricing policies are depicted in Table 13. The first line is a restatement of the uniform pricing scheme from Table 10 that earns a profit of $734.80. We treat this as the baseline profit for comparing the store-level pricing policies that tailor the private label price to the demand conditions of each individual store. With no cross-shopping considerations, profits increase by 9.6% to $805.47. Adding in the cross-shopping constraints, the pricing policy still achieves a 6.2% increase to $780.42 over the baseline profits. Thus, even with the cross-shopping constraints, customizing prices still allows the retailer to exploit demand heterogeneity to a meaningful degree. These constrained targeted prices resemble in many respects the idea of zone pricing (Meza and Sudhir 2010; Hruschka 2007).

## *CONCLUSION*

This paper introduces a new approach to forecasting sales of products with new-to-market attributes that integrates choice-based conjoint data with repeated purchase data for a dense consumer panel. We demonstrate that our model improves predictions over conjoint alone and a strong benchmark model in a series of in-sample and hold-out predictive validity tests on actual purchases made months later when the new product was launched. Our data also allow us to evaluate the biases that limit the predictive validity of the competing models, and we find while both selection and context biases exist, after controlling for observable heterogeneity, the contextual biases are much larger. We then apply our model to inform decisions about the optimal product line to launch and the optimal store-level prices to charge. This exercise demonstrates the potential value of the approach.

We suggest that our method could be applied in a great many settings. Of course, any retailer can use this approach to price new product launches in their own stores by integrating the conjoint data with loyalty card data. Because our method does not require purchases for the same individuals, brand managers can apply this to household scanner data available, for example, from Nielsen. This opens the method to virtually any grocery product launch. Further, although we focus on an application were the $i$ subscript is for individuals, one can recast our problem at the micro-segment level. In this context, our approach can be applied to launches for online retailers like Amazon or Sephora, who can obtain panels of individuals with similar characteristics. In fact, any time purchase histories are available for a representative sample of individuals, businesses, or micro-segments, our approach can be applied.

This research comes with some limitations that point to potential avenues for future research. First, we demonstrate our approach for only one application context. While our model is "general" in the sense of nesting settings without conjoint preference distortions, different distortions might be more difficult to address than those uncovered here. Further testing of our approach is left to future research, and we caution against overly generalizing from this one application. However, noting that it has not been historically easy to obtain repeated actual purchase settings for predictive validation, as evidenced by the limited demonstrations in the literature, our predictive validation is relatively strong evidence. Second, as is often the case for surveys of customer groups, in our setting we were not able to use the kind of incentives that have been shown to further improve the validity of conjoint estimates (Ding et al., 2005). Future research could evaluate the degree to which incentives can reduce the preference distortions arising from context effects. Third, for our sample sizes, models, and available computational resources,we were limited in how precisely we could identify the preference shifters that improve the linking between actual and conjoint choices. Ideally, one would apply a LASSO-type approach with cross-validation to empirically select the preference shifters. Future research could extend the ideas here to identify a computationally efficient way to incorporate a LASSO-type approach.

## ***References***

Adamowicz, W., J. Louviere, and M. Williams (1994). Combining revealed and stated preference methods for valuing environmental amenities. *Journal of environmental economics and management 26*(3), 271–292.

Ailawadi, K. L., K. Pauwels, and J.-B. E. Steenkamp (2008). Private-label use and store loyalty. *Journal of Marketing 72*(6), 19–30.

Allenby, G., G. Fennell, J. Huber, T. Eagle, T. Gilbride, D. Horsky, J. Kim, P. Lenk, R. Johnson, E. Ofek, et al. (2005). Adjusting choice models to better predict market behavior. *Marketing Letters 16*(3), 197–208.

Ben-Akiva, M., D. McFadden, and K. Train (2015). Foundations of stated preference elicitation. Technical report, working paper, Department of Economics, University of California, Berkeley, http://eml. berkeley. edu/˜ train/foundations. pdf.

Ben-Akiva, M. and T. Morikawa (1990). Estimation of switching models from revealed preferences and stated intentions. *Transportation Research A 24*(6), 489–495.

Brooks, K. and J. L. Lusk (2010). Stated and revealed preferences for organic and cloned milk: combining choice experiment and scanner data. *American Journal of Agricultural Economics*, aaq054.

Brownstone, D., D. S. Bunch, and K. Train (2000). Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles. *Transportation Research Part B: Methodological 34*(5), 315–338.

Chang, J. B., J. L. Lusk, and F. B. Norwood (2009). How closely do hypothetical surveys and laboratory experiments predict field behavior? *American Journal of Agricultural Economics 91*(2), 518–534.

Cherchi, E. and J. de Dios Ortúzar (2002). Mixed rp/sp models incorporating interaction effects. *Transportation 29*(4), 371–395.

Chintagunta, P. K., A. Bonfrer, and I. Song (2002). Investigating the effects of store-brand introduction on retailer demand and pricing behavior. *Management Science 48*(10), 1242–1267.

DeGroot, M. (1970). *Optimal Statistical Decisions.* McGraw-Hill.

Ding, M. (2007). An incentive-aligned mechanism for conjoint analysis. *Journal of Marketing Research 44*(2), 214–223.

Ding, M., R. Grewal, and J. Liechty (2005). Incentive-aligned conjoint analysis. *Journal of marketing research 42*(1), 67–82.

Ding, M. and J. Huber (2009). When is hypothetical bias a problem in choice tasks, and what can we do about it? In *Proceedings of the Sawtooth Software Conference*, pp. 263–272.

Dong, S., M. Ding, and J. Huber (2010). A simple mechanism to incentive-align conjoint experiments. *International Journal of Research in Marketing 27*(1), 25–32.

Dubé, J.-P., G. J. Hitsch, and P. E. Rossi (2010). State dependence and alternative explanations for consumer inertia. *The RAND Journal of Economics 41*(3), 417–445.

Feit, E. M., M. A. Beltramo, and F. M. Feinberg (2010). Reality check: Combining choice experiments with market data to estimate the importance of product attributes. *Management Science 56*(5), 785–800.

Feurstein, M., M. Natter, and L. Kehl (1999). Forecasting scanner data by choice-based conjoint models. In *Sawtooth Software Conference Proceedings.*

Geyskens, I., K. Gielens, and E. Gijsbrechts (2010). Proliferating private-label portfolios: How introducing economy and premium private labels influences brand choice. *Journal of Marketing Research 47*(5), 791–807.

Green, P. E. and V. R. Rao (1971). Conjoint measurement for quantifying judgmental data. *Journal of Marketing research*, 355–363.

Green, P. E. and V. Srinivasan (1990). Conjoint analysis in marketing: new developments with implications for research and practice. *The Journal of Marketing*, 3–19.

Hensher, D., J. Louviere, and J. Swait (1998). Combining sources of preference data. *Journal of Econometrics 89*(1), 197–221.

Horsky, D. and P. Nelson (1992). New brand positioning and pricing in an oligopolistic market. *Marketing Science 11*(2), 133–153.

Hruschka, H. (2007). Clusterwise pricing in stores of a retail chain. *OR Spectrum 29*(4), 579–595.

Islam, T., J. J. Louviere, and P. F. Burke (2007). Modeling the effects of including/excluding attributes in choice experiments on systematic and random components. *International Journal of Research in Marketing 24*(4), 289–300.

Louviere, J., D. Hensher, and J. Swait (2007). Conjoint preference elicitation methods in the broader context of random utility theory preference elicitation methods. In A. H. Anders Gustafsson and F. Huber (Eds.), *Conjoint Measurement: Methods and Applications* (4th ed.)., Chapter 10, pp. 167–198. Berlin: Springer.

Louviere, J. J., D. A. Hensher, and J. D. Swait (2000). *Stated choice methods: analysis and applications.* Cambridge University Press.

Manski, C. F. and S. R. Lerman (1977). The estimation of choice probabilities from choice based samples. *Econometrica: Journal of the Econometric Society*, 1977–1988.

Meza, S. and K. Sudhir (2010). Do private labels increase retailer bargaining power? *Quantitative Marketing and Economics 8*(3), 333–363.

Orme, B. K. (2014). *Getting started with conjoint analysis: strategies for product design and pricing research.* Research Publishers.

Orme, B. K., M. I. Alpert, and E. Christensen (1997). Assessing the validity of conjoint analysis–continued. In *Sawtooth Software Conference Proceedings*, pp. 209–226.

Orme, B. K. and M. Heft (1999). Predicting actual sales with cbc: How capturing heterogeneity improves results. In *Sawtooth Software Conference Proceedings*, pp. 183–199.

Ozer, M. and W. Kamakura (2007). A multi-trait multi-method validity test of partwork estimates. In A. H. Anders Gustafsson and F. Huber (Eds.), *Conjoint Measurement: Methods and Applications* (4th ed.)., Chapter 9, pp. 145–166. Berlin: Springer.

Rogers, G. and T. Renken (2003). Validation and calibration of choice-based conjoint for pricing research. In *Sawtooth Software Conference Proceedings*, pp. 209–215.

Swait, J. and R. L. Andrews (2003). Enriching scanner panel models with choice experiments. *Marketing Science 22*(4), 442–460.

Swait, J. and J. Louviere (1993). The role of the scale parameter in the estimation and comparison of multinomial logit models. *Journal of marketing research*, 305–314.

Toubia, O., D. I. Simester, J. R. Hauser, and E. Dahan (2003). Fast polyhedral adaptive conjoint estimation. *Marketing Science 22*(3), 273–303.

Trope, Y. and N. Liberman (2010). Construal-level theory of psychological distance. *Psychological review 117*(2), 440.

Whitehead, J. C., S. K. Pattanayak, G. L. Van Houtven, and B. R. Gelso (2008). Combining revealed and stated preference data to estimate the nonmarket value of ecological services: an assessment of the state of the science. *Journal of Economic Surveys 22*(5), 872–908.

Wittink, D. (2003). Comment on arenoe and rogers/renken. In *Sawtooth Software Conference 2003 Proceedings*, pp. 233–236.

Wittink, D. R. (2000). Predictive validation of conjoint analysis. In *Sawtooth Software Conference Proceedings, Sequim, WA*. Citeseer.

Wittink, D. R. and P. Cattin (1989). Commercial use of conjoint analysis: An update. *The Journal of Marketing*, 91–96.

***TABLES***

| Brand | Total Brand Shares (%) | | Inside Good Shares (%) | |
|---|---|---|---|---|
| | $RP_{rs}$ | $RP_{ss}$ | $RP_{rs}$ | $RP_{ss}$ |
| Brown Cow | 1.68 | 6.89 | 8.34 | 10.24 |
| Chobani | 12.62 | 43.15 | 62.66 | 64.13 |
| Dannon | 2.78 | 7.06 | 13.80 | 10.49 |
| Fage | 1.70 | 6.31 | 8.44 | 9.38 |
| Oikos | 1.35 | 3.88 | 6.70 | 5.77 |
| Outside Good | 79.86 | 32.72 | | |

Table 1: Brand Shares for $RP_{rs}$ and $RP_{ss}$ during the estimation period

| Summary Stats | Random Sample ($RP_{rs}$) | Random Estimation Sample | Survey Sample ($RP_{ss}$) | Survey Sample ($SP_{ss}$) |
|---|---|---|---|---|
| No. of individuals | 4132 | 510 | 504* | 510 |
| Period Starting | Feb 6, 2011 | Feb 6, 2011 | Feb 6, 2011 | May 20, 2011 |
| Mean # of days since last Greek yogurt purchase | 386.6 | 364.6 | 165.7 | 25.9 |
| Mean # of visits with yogurt purchase in past 90 days | 0.4 | 0.5 | 3.2 | 8.0 |
| Mean sales in Greek yogurt per store visit in past 90 days ($) | 0.4 | 0.4 | 2.0 | 3.9 |
| % Households $< 75k$ | 51.5 | 52.4 | 38.0 | 38.0 |
| % Households $\leq 3$ Members | 59.2 | 58.8 | 58.3 | 58.3 |

* $RP_{ss}$ sample sizes differs from $SP_{ss}$ due to six individuals not visiting any store during a period.

Table 2: Demographics and RFM variables for random and survey samples

| Summary Stats | Random Estimation Sample ($RP_{rs}$) | Survey Estimation Sample ($RP_{ss}$) |
|---|---|---|
| No. of individuals | 510 | 504 |
| No. of choice occasions | 28731 | 29640 |
| No. of products | 15 | 15 |
| % product availability | 97.75 | 97.26 |
| Mean (S.E.) | | |
| Price | $1.22 (0.41) | $1.20 (0.42) |
| State dependence | 0.05 (0.21) | 0.17 (0.21) |

Table 3: Descriptive Statistics for $RP_{rs}$ and $RP_{ss}$ choice sets

| Attributes | $RP_{rs}$ Data | | $SP_{ss}$ Data | |
|---|---|---|---|---|
| | $\Delta_0$ (S.E.) | $\sigma$ (S.E.) | $\Delta_0$ (S.E.) | $\sigma$ (S.E.) |
| Inside Good | **-8.68 (0.19)** | **6.31 (0.11)** | **2.11 (0.25)** | **3.18 (0.22)** |
| Brown Cow | **0.35 (0.14)** | **2.01 (0.11)** | | |
| Chobani | **3.91 (0.12)** | **3.65 (0.09)** | **1.18 (0.13)** | **1.56 (0.14)** |
| Fage | **0.96 (0.33)** | **5.41 (0.28)** | 0.40 (0.40) | **2.76 (0.17)** |
| Oikos | **-1.12 (0.29)** | **6.43 (0.21)** | **0.60 (0.19)** | **1.58 (0.16)** |
| Private Label | | | **1.87 (0.36)** | **1.88 (0.13)** |
| Zero-Fat | **0.72 (0.07)** | **1.49 (0.06)** | **1.10 (0.10)** | **1.88 (0.10)** |
| Fruit | **1.94 (0.09)** | **3.79 (0.10)** | **0.31 (0.11)** | **2.32 (0.11)** |
| Organic | | | 0.13 (0.12) | **1.17 (0.12)** |
| All Natural | | | **-0.17 (0.07)** | **0.39 (0.12)** |
| Probiotic | | | **0.12 (0.05)** | 0.12 (0.09) |
| Vitamins | | | 0.07 (0.08) | **0.44 (0.13)** |
| Omega-3 | | | -0.07 (0.08) | **0.79 (0.10)** |
| Fiber | | | -0.03 (0.09) | **0.86 (0.10)** |
| SBS Cup | **-1.36 (0.42)** | **4.02 (0.37)** | **-1.37 (0.36)** | **1.86 (0.18)** |
| SBS Cup Incl. | | | **-1.72 (0.12)** | **1.84 (0.12)** |
| Normal Cup Incl. Top | | | **-1.43 (0.33)** | **0.62 (0.21)** |
| Price | **-10.08 (0.20)** | **7.41 (0.14)** | **-2.64 (0.20)** | **1.25 (0.14)** |
| State Dependence | **0.84 (0.06)** | **1.64 (0.06)** | **2.52 (0.15)** | **1.57 (0.16)** |
| No. of individuals | 510 | | 510 | |
| No. of alternatives | 15 | | 5 | |
| No. of choice occasions | 28731 | | 6208 | |
| Log-Likelihood | -10743.07 | | -6278.63 | |
| AIC | 21626.14 | | 12827.26 | |
| BIC | 22204.74 | | 13675.69 | |

Bold indicates 95% significance

Table 4: Standalone estimation of $RP_{rs}$ and $SP_{ss}$ data

| Attributes | BENCH Estimation | | CPC Estimation | |
|---|---|---|---|---|
| | $\Delta_0^{SP_{ss}}$ (S.E.) | $\sigma^{SP_{ss}}$ (S.E.) | $\Delta_0^{SP_{ss}}$ (S.E.) | $\sigma^{SP_{ss}}$ (S.E.) |
| Inside Good | **-1.81 (0.09)** | **6.03 (0.20)** | **2.64 (0.22)** | **3.04 (0.08)** |
| Brown Cow | **-0.36 (0.10)** | **1.29 (0.08)** | -0.31 (0.19) | **1.10 (0.12)** |
| Chobani | **2.01 (0.08)** | **1.38 (0.05)** | **1.51 (0.10)** | **2.03 (0.06)** |
| Fage | **0.48 (0.14)** | **2.40 (0.11)** | **-0.43 (0.17)** | **2.83 (0.12)** |
| Oikos | **0.96 (0.13)** | **0.62 (0.07)** | **0.72 (0.13)** | **1.51 (0.10)** |
| Private Label | **1.79 (0.19)** | **1.57 (0.13)** | **1.12 (0.26)** | **1.96 (0.13)** |
| Zero-Fat | **0.82 (0.04)** | **1.21 (0.05)** | **0.68 (0.05)** | **1.16 (0.04)** |
| Fruit | **0.67 (0.04)** | **1.65 (0.06)** | **0.25 (0.11)** | **1.63 (0.12)** |
| Organic | -0.08 (0.12) | **1.62 (0.08)** | 0.11 (0.12) | **1.22 (0.07)** |
| All Natural | -0.04 (0.06) | **0.40 (0.03)** | -0.03 (0.06) | **0.18 (0.02)** |
| Probiotic | **0.18 (0.05)** | 0.01 (0.08) | **0.12 (0.05)** | 0.04 (0.10) |
| Vitamins | 0.08 (0.07) | 0.24 (0.13) | 0.06 (0.07) | 0.18 (0.18) |
| Omega-3 | 0.01 (0.07) | 0.22 (0.17) | -0.08 (0.08) | **0.60 (0.12)** |
| Fiber | 0.05 (0.07) | **0.36 (0.12)** | 0.02 (0.08) | **0.59 (0.15)** |
| SBS Cup | **-1.30 (0.17)** | **1.60 (0.10)** | **-0.66 (0.34)** | **1.69 (0.24)** |
| SBS Cup Incl. | **-1.47 (0.10)** | **1.63 (0.13)** | **-1.65 (0.12)** | **1.71 (0.11)** |
| Normal Cup Incl. Top | **-1.21 (0.17)** | **1.00 (0.13)** | **-0.65 (0.20)** | 0.21 (0.39) |
| Price | **-2.97 (0.12)** | **3.34 (0.11)** | **-2.26 (0.19)** | **2.62 (0.07)** |
| State Dependence | **0.97 (0.05)** | **1.30 (0.05)** | **2.20 (0.12)** | **0.97 (0.05)** |
| Scaling Factor $\lambda$ | **1.80 (0.06)** | | **1.91 (0.08)** | |
| Additive Shifter $\mu_0$ for Inside Good | | | **-7.20 (0.24)** | |
| Additive Shifter $\mu_0$ for Fruit | | | **0.60 (0.11)** | |
| Additive Shifter $\mu_0$ for Price | | | **-1.64 (0.33)** | |
| Additive Shifter $\mu_0$ for State Dep. | | | **-1.88 (0.15)** | |
| No. of $RP_{rs}$ Choice occasions | 28731 | | 28731 | |
| No. of $SP_{ss}$ Choice occasions | 6208 | | 6208 | |
| $RP_{rs}$ Log-Likelihood | -11046.74 | | -10769.83 | |
| $SP_{ss}$ Log-Likelihood | -6774.42 | | -6486.15 | |
| Total Log-Likelihood | -17821.15 | | -17255.98 | |
| AIC | 35910.30 | | 34787.95 | |
| BIC | 37044.12 | | 35955.62 | |

Bold indicates 95% significance

Table 5: Comparison of $BENCH$ and $CPC$ Estimation

| Model | Estimation Period | | Forecast Period 1 After Brown Cow Exit | | Forecast Period 2 After Store Label Entry | |
|---|---|---|---|---|---|---|
| | $RP_{rs}$ Est. | $RP_{rs}$ Hold. | $RP_{rs}$ Est. | $RP_{rs}$ Hold. | $RP_{rs}$ Est. | $RP_{rs}$ Hold. |
| # Indiv. | 510 | 3622 | 510 | 2223 | 510 | 2263 |
| # Choices | 28731 | 134693 | 7717 | 28766 | 7389 | 28145 |
| $SP_{ss}$ Only | NA | NA | -4657.7 | -16869.9 | -4534.6 | -18388.6 |
| $RP_{rs}$ Only | -10743.1 | -54417.6 | -3132.1 | -10997.1 | NA | NA |
| BENCH | -11046.7 | -54947.6 | -3329.6 | -11549.8 | -3203.3 | -12610.1 |
| CPC | -10769.8 | -53183.2 | -3138.9 | -10839.2 | -3088.7 | -11976.0 |

Table 6: Comparison of predictive log-likelihoods on actual purchases for estimation and holdout samples across models and time periods

| Model | Estimation Period | | Forecast Period 1 After Brown Cow Exit | | Forecast Period 2 After Store Label Entry | |
|---|---|---|---|---|---|---|
| | $RP_{rs}$ Est. | $RP_{rs}$ Hold. | $RP_{rs}$ Est. | $RP_{rs}$ Hold. | $RP_{rs}$ Est. | $RP_{rs}$ Hold. |
| # Indiv. | 510 | 3622 | 510 | 2223 | 510 | 2263 |
| # Choices | 28731 | 134693 | 7717 | 28766 | 7389 | 28145 |
| $SP_{ss}$ Only | NA | NA | 0.8307 | 0.8530 | 0.8571 | 0.8633 |
| $RP_{rs}$ Only | 0.3789 | 0.4061 | 0.3968 | 0.4069 | NA | NA |
| BENCH | 0.5735 | 0.5779 | 0.5894 | 0.6131 | 0.6336 | 0.6320 |
| CPC | 0.3823 | 0.4035 | 0.4124 | 0.4263 | 0.4555 | 0.4540 |

Table 7: Comparison of RMSE in individual demand predictions versus actual purchases for estimation and holdout samples across models and time periods

| Model | Estimation Period | | Forecast Period 1 After Brown Cow Exit | | Forecast Period 2 After Store Label Entry | |
|---|---|---|---|---|---|---|
| | $RP_{rs}$ Est. | $RP_{rs}$ Hold. | $RP_{rs}$ Est. | $RP_{rs}$ Hold. | $RP_{rs}$ Est. | $RP_{rs}$ Hold. |
| # Indiv. | 510 | 3622 | 510 | 2223 | 510 | 2263 |
| # Choices | 28731 | 134693 | 7717 | 28766 | 7389 | 28145 |
| $SP_{ss}$ Only | NA | NA | 0.7394 | 0.7365 | 0.7476 | 0.7399 |
| $RP_{rs}$ Only | 0.2066 | 0.1465 | 0.1987 | 0.1490 | NA | NA |
| BENCH | 0.4581 | 0.4169 | 0.4260 | 0.4021 | 0.4493 | 0.4195 |
| CPC | 0.2118 | 0.1467 | 0.2061 | 0.1461 | 0.2315 | 0.1835 |

Table 8: Comparison of RMSE in store demand predictions on actual purchases for estimation and holdout samples across models and time periods

| Attributes | CPC Estimation with Dem. & RFM Mean (S.E.) | CPC Estimation with Dem. & RFM Diff. (S.E.)[1] | CPC Estimation with Dem. Only Mean (S.E.) | CPC Estimation with Dem. Only Diff. (S.E.)[1] | Diff[2] in-Diff |
|---|---|---|---|---|---|
| Scaling Factor $\lambda^{RP_{ss}}$ | **1.83 (0.06)** | **0.17 (0.04)** | **1.74 (0.06)** | **0.26 (0.04)** | 0.09 (0.06) |
| Scaling Factor $\lambda^{RP_{rs}}$ | **2.00 (0.07)** | | **2.00 (0.08)** | | |
| $\mu_0^{RP_{ss}}$ for Inside Good | **-5.59 (0.27)** | **-1.18 (0.07)** | **-5.01 (0.24)** | **-1.60 (0.08)** | **-0.42 (0.11)** |
| $\mu_0^{RP_{rs}}$ for Inside Good | **-6.77 (0.29)** | | **-6.61 (0.28)** | | |
| $\mu_0^{RP_{ss}}$ for Fruit | **0.64 (0.09)** | **-0.14 (0.04)** | **0.56 (0.09)** | 0.03 (0.04) | **0.17 (0.06)** |
| $\mu_0^{RP_{rs}}$ for Fruit | **0.51 (0.09)** | | **0.59 (0.09)** | | |
| $\mu_0^{RP_{ss}}$ for Price | **-1.05 (0.22)** | **-0.25 (0.07)** | **-1.17 (0.20)** | **-1.25 (0.09)** | **-1.00 (0.12)** |
| $\mu_0^{RP_{rs}}$ for Price | **-1.30 (0.22)** | | **-2.41 (0.23)** | | |
| $\mu_0^{RP_{ss}}$ for State Dep. | **-1.41 (0.13)** | **-0.43 (0.04)** | **-1.30 (0.12)** | **-0.23 (0.04)** | **0.20 (0.06)** |
| $\mu_0^{RP_{rs}}$ for State Dep. | **-1.84 (0.13)** | | **-1.53 (0.12)** | | |
| $SP_{ss}$ Log-Likelihood | -6576.63 | | -6638.90 | | |
| $RP_{ss}$ Log-Likelihood | -30736.47 | | -31029.48 | | |
| $RP_{rs}$ Log-Likelihood | -10923.11 | | -11012.24 | | |
| Total Log-Likelihood | -48236.21 | | -48680.62 | | |
| # $SP_{ss}$ Choice occasions | 6208 | | | | |
| # $RP_{ss}$ Choice occasions | 29640 | | | | |
| # $RP_{rs}$ Choice occasions | 28731 | | | | |

Bold indicates 95% significance

[1] Differences in $\lambda$ and $\mu$ between $RP_{rs}$ and $RP_{ss}$ datasets. Standard errors calculated by the Delta Method.

[2] Comparing ($RP_{rs}$-$RP_{ss}$) differences across models via unpaired t-test.

Table 9: Comparing Contextual and Selection Biases

| Attributes | Unconstrained Maximization Exp. Profit* ($) | Unconstrained Maximization Price ($) | Constrained Maximization Exp. Profit*+ ($) | Constrained Maximization Price ($) |
|---|---|---|---|---|
| *Cheap Attributes:* | | | | |
| Baseline | 550.55 | 1.07 | 527.26 | 0.89 |
| Zero Fat | 794.17 | 1.14 | 734.80 | 0.89 |
| *Expensive Attributes:* | | | | |
| Organic | 581.68 | 1.08 | 554.91 | 0.89 |
| All-Natural | 592.67 | 1.10 | 560.29 | 0.89 |
| Probiotic | 608.77 | 1.11 | 574.17 | 0.89 |
| Vitamins | 509.73 | 1.02 | 497.30 | 0.89 |
| Omega-3 | 577.57 | 1.09 | 548.93 | 0.89 |
| Fiber | 621.75 | 1.14 | 580.45 | 0.89 |
| Side-by-Side Cup | 417.49 | 0.98 | 411.95 | 0.89 |
| Side-by-Side Cup with Inclusions | 444.95 | 1.04 | 430.64 | 0.89 |
| Normal Cup with Inclusions on Top | 366.56 | 0.94 | 365.13 | 0.89 |

[*] Expected profits computed with $MAP$ estimates for existing attributes and integrating over 10000 random normal draws for NTMA attributes

[+] Upper limit for prices set at $0.89

Table 10: Expected profits for different potential products at $c_{PL} = \$0.40$

| Marginal Cost ($) | Constr. Maximum Profit* ($) | Constrained Price+ ($) |
|---|---|---|
| *Basic Product Line* | | |
| 0.40 | 734.80 | 0.89 |
| *Composite Product Line* | | |
| 0.45 | 940.56 | 1.34 |
| 0.50 | 889.77 | 1.41 |
| 0.55 | 843.25 | 1.50 |
| 0.60 | 801.77 | 1.65 |
| 0.65 | 764.91 | 1.69 |
| 0.70 | 728.13 | 1.69 |
| 0.75 | 691.36 | 1.69 |

* Expected profits computed with $MAP$ estimates for existing attributes and integrating over 10000 random normal draws for NTMA attributes

+ Prices for basic line constrained at $0.89 and composite line at $1.69

Table 11: Comparing profitability of basic and composite product-lines

| | | | | |
|---|---|---|---|---|
| No. of stores visited | 1 | 2 | 3 | 4 |
| No. of individuals | 2410 | 310 | 42 | 5 |

Table 12: Distribution of individuals and no. of stores visited

| Pricing Strategy | Brand pricing constraints | Cross-Shopping constraints | Max. Profit ($)* | Maximizing+ Price(s) ($) |
|---|---|---|---|---|
| Constrained Uniform Pricing | ✓ | ✓ | 734.80 | 0.89 |
| Store pricing with brand pricing constraints | ✓ | | 805.47 | min - 0.72, med - 1.06, max - 1.19 |
| Store Pricing with brand and cross-shopping constraints | ✓ | ✓ | 780.42 | min - 0.87, med - 0.98, max - 1.19 |

* Expected profits computed with $MAP$ estimates for existing attributes and integrating over 10000 random normal draws for NTMA attributes

+ For store-level prices, min, max and median of optimized prices across stores are reported

Table 13: Constrained Pricing profits

## *FIGURES*

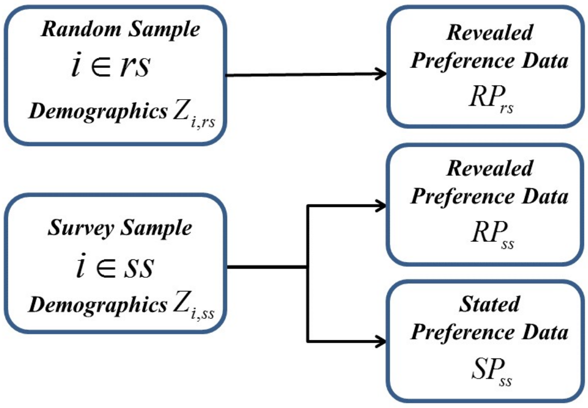


Figure 1: Relationship between Samples and Datasets

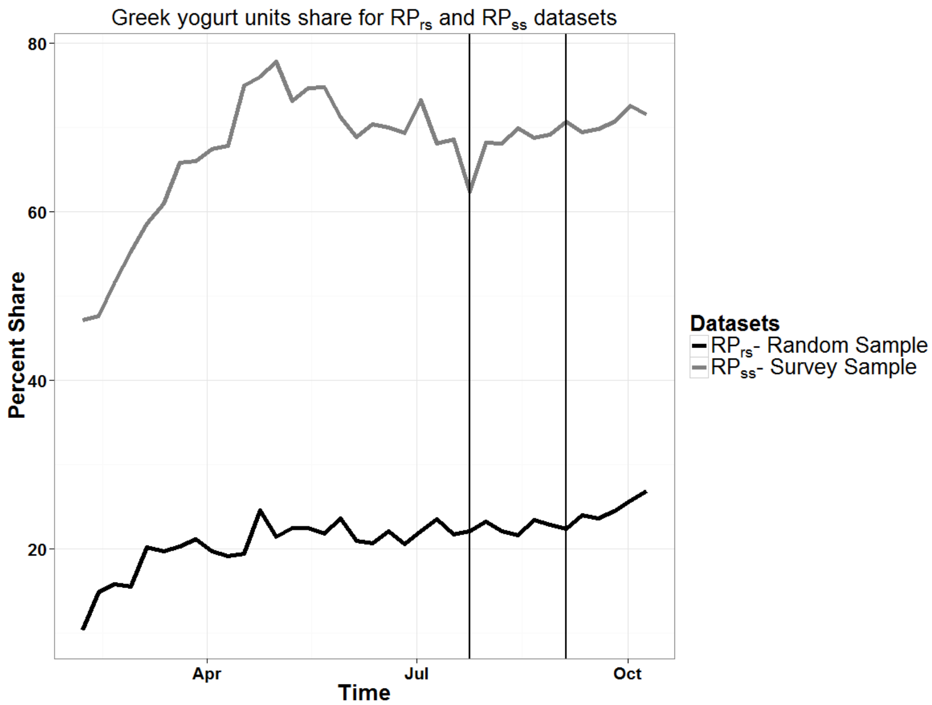


Figure 2: Evolution of Greek yogurt shares across $RP_{rs}$ and $RP_{ss}$ datasets

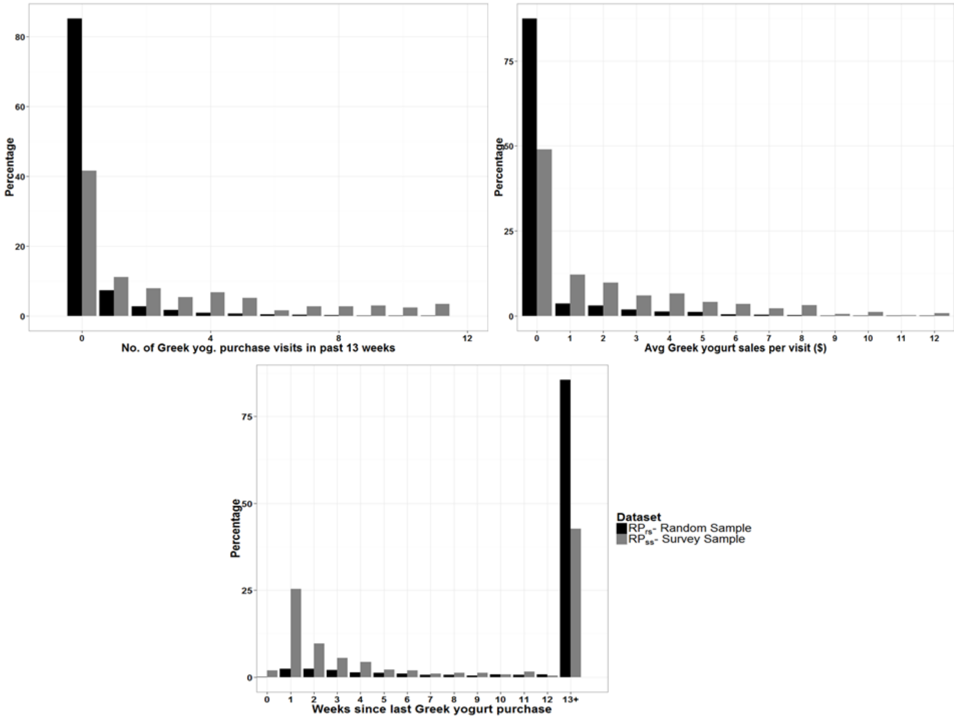


Figure 3: Comparing RFM variables across $RP_{rs}$ and $RP_{ss}$ datasets

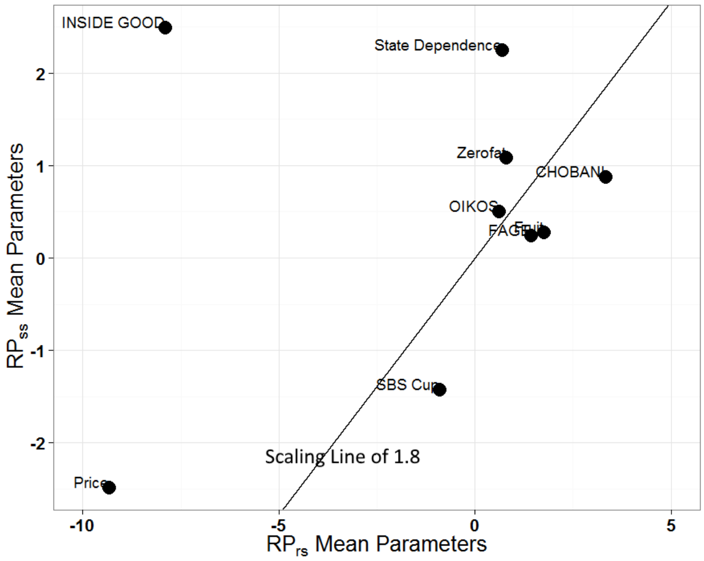


Figure 4: Comparing common coefficients across models

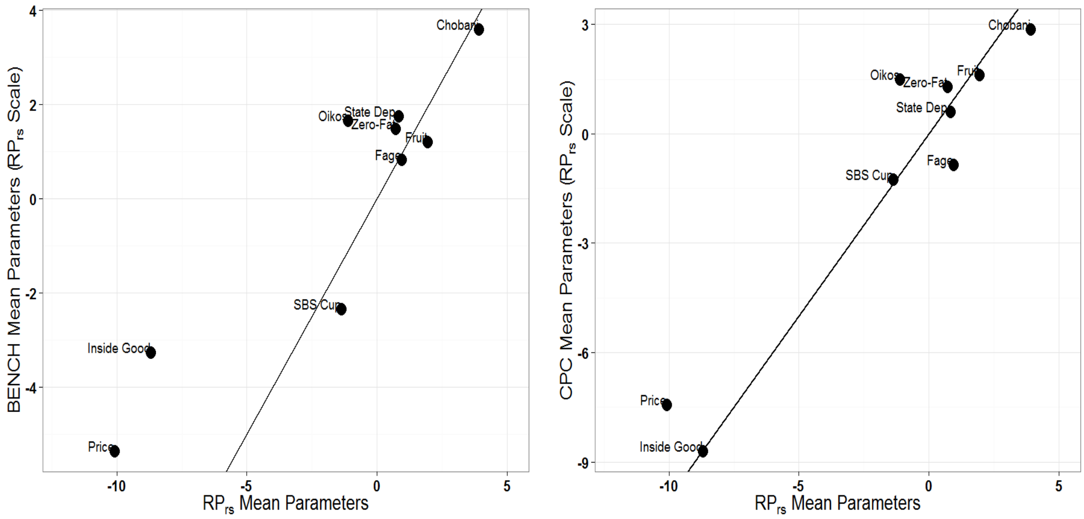


Figure 5: Comparing common coefficients across models

## *WEB APPENDIX*

*Web Appendix A: Jacobian And Hessian For SML Function*

As noted in the main text, the aggregate parameters of interest are $\Theta = \{\Delta_0^{SP_{ss}}, \Delta^{SP_{ss}}, \Sigma^{SP_{ss}}, \lambda^{RP_{ss}}, \mu_0^{RP_{ss}}, \lambda^{RP_{rs}}, \mu_0^{RP_{rs}}\}$. The log-likelihood function for the $RP_{rs} - SP_{ss}$ SMLE can be written as[10]

$$
\begin{aligned}
LL(\Theta) &= \sum_{i \in SP_{ss}} \log \mathcal{L}_i^{SP_{ss}}(\Theta) + \sum_{i'' \in RP_{rs}} \log \mathcal{L}_{i''}^{RP_{rs}}(\Theta) \\
LL(\Theta) &= \sum_{i \in SP_{ss}} \log \int \prod_{t=1}^{T_i} P_{i[j]t}^{SP_{ss}}(\Theta, u_i) f(u_i) du_i \\
&\quad + \sum_{i'' \in RP_{rs}} \log \int \prod_{t=1}^{T_{i''}} P_{i''[j]t}^{RP_{rs}}(\Theta, u_{i''}) f(u_{i''}) du_{i''} \\
LL(\Theta) &\simeq \sum_{i \in SP_{ss}} \log \big( \frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T_i} P_{i[j]t}^{SP_{ss}}(\Theta, u_i^r) \big) \\
&\quad + \sum_{i'' \in RP_{rs}} \log \big( \frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T_{i''}} P_{i''[j]t}^{RP_{rs}}(\Theta, u_{i''}^r) \big)
\end{aligned}
$$

where $u_i^r$, $u_{i''}^r$ are $R$ iid draws from the probability density $f(\cdot)$. Since the log-likelihoods are similar in structure barring a few differences (additive and multiplicative scaling constants for the $RP_{rs}$ data), we will write down the jacobians and hessians for the $SP_{ss}$ log-likelihood function and then highlight the differences for the $RP_{rs}$ case.

$$
\begin{aligned}
LL^{SP_{ss}}(\Theta) &\simeq \sum_{i \in SP_{ss}} \log \underbrace{\big( \frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T_i} P_{i[j]t}^{SP_{ss}}(\Theta, u_i^r) \big)}_{\mathcal{L}_i^{SP_{ss}}} \\
\frac{dLL^{SP_{ss}}}{d\Theta} &\simeq \sum_{i \in SP_{ss}} \frac{1}{\mathcal{L}_i^{SP_{ss}}} \frac{1}{R} \sum_{r=1}^{R} \frac{d}{d\Theta} \big[ \prod_{t=1}^{T_i} P_{i[j]t}^{SP_{ss}}(\Theta, u_i^r) \big] \\
&\simeq \sum_{i \in SP_{ss}} \frac{1}{\mathcal{L}_i^{SP_{ss}}} \frac{1}{R} \sum_{r=1}^{R} \big[ \prod_{t=1}^{T_i} P_{i[j]t}^{SP_{ss}}(\Theta, u_i^r) \big] \left( \sum_{t=1}^{T_i} \frac{1}{P_{i[j]t}^{SP_{ss}}} \frac{d}{d\Theta} P_{i[j]t}^{SP_{ss}}(\Theta, u_i^r) \right)
\end{aligned}
$$

[10]The authors can be contacted for similar derivations for the log-likelihoods, jacobian and hessian functions for the $RP_{rs}$-$RP_{ss}$-$SP_{ss}$ Estimation.

Define $l_{itr} = P^{SP_{ss}}_{i[j]t}(\Theta, u_i^r)$. Then $l_{ir} = \prod_{t=1}^{T_i} l_{itr}$ and $\mathcal{L}_i^{SP_{ss}}(\Theta) = \frac{1}{R}\sum_{r=1}^{R} l_{ir}$.

(19)
$$\frac{dLL^{SP_{ss}}}{d\Theta} \simeq \sum_{i \in SP_{ss}} \sum_{r=1}^{R} \frac{l_{ir}}{\sum_{r=1}^{R} l_{ir}} (\sum_{t=1}^{T_i} \frac{1}{l_{itr}} \frac{dl_{itr}}{d\Theta})$$

where

$$l_{itr} = \frac{\exp(X_{ijt}\beta_{ir})}{1 + \sum_{k=1}^{J}\exp(X_{ikt}\beta_{ir})}$$

(20)
$$\frac{dl_{itr}}{d\Theta} = l_{itr}(X_{ijt} - \sum_{k=1}^{J} l_{iktr} X_{ikt}) \frac{d\beta_{ir}}{d\Theta}$$

Here $j$ is the product chosen, and $l_{iktr}$ is the logit probability of choosing product $k$ in trip $t$ with unobserved heterogeneity component $u_i^r$. Substituting (20) back in (19),

(21)
$$\begin{aligned} \frac{dLL}{d\Theta} &\simeq \sum_{i \in SP_{ss}} \sum_{r=1}^{R} \underbrace{\frac{l_{ir}}{\sum_{r=1}^{R} l_{ir}} \left( \sum_{t=1}^{T_i} \overbrace{(X_{ijt} - \sum_{k=1}^{J} l_{iktr} X_{ikt})}^{1XK} \quad \overbrace{\frac{d\beta_{ir}}{d\Theta}}^{KX(nz+1)K+5} \right)}_{J_{ir}} \\ &\simeq \sum_{i \in SP_{ss}} \sum_{r=1}^{R} \frac{1}{\sum_{r=1}^{R} l_{ir}} \frac{dl_{ir}}{d\Theta} = \sum_{i \in SP_{ss}} \frac{dLL_i}{d\Theta} \end{aligned}$$

From (21), we can see that the structure of the jacobian function is fairly generic. The $RP_{rs}$ and $SP_{ss}$ jacobians differ in their respective $X$ matrices, and the $\frac{\partial \beta_{ir}}{\partial \Theta}$ matrix. Recall that

$$\begin{aligned} \beta_{ir}^{SP_{ss}} &= \Delta_0^{SP_{ss}} + Z_{i,ss}\Delta^{SP_{ss}} + \Sigma^{SP_{ss}}\nu_i^r \\ \beta_{i''r}^{RP_{rs}} &= \lambda^{RP_{rs}}(\Delta_0^{SP_{ss}} + \mu_0^{RP_{rs}} + Z_{i,rs}\Delta^{SP_{ss}} + \Sigma^{SP_{ss}} u_{i''}^r) \end{aligned}$$

Hence,

$$\frac{\partial \beta_{ir}^{SP_{ss}}}{\partial \Delta_0} = \mathbb{I}_K \qquad \frac{\partial \beta_{i''r}^{RP_{rs}}}{\partial \Delta_0} = \lambda^{RP_{rs}} \mathbb{I}_K$$

$$\frac{\partial \beta_{ir}^{SP_{ss}}}{\partial \Delta^{SP_{ss}}} = Z_{i,ss} \qquad \frac{\partial \beta_{i''r}^{RP_{rs}}}{\partial \Delta^{SP_{ss}}} = \lambda^{RP_{rs}} Z_{i,rs}$$

$$\frac{\partial \beta_{ir}^{SP_{ss}}}{\partial \Sigma} = \mathbb{I}_K u_i^r \qquad \frac{\partial \beta_{i''r}^{RP_{rs}}}{\partial \Sigma} = \lambda^{RP_{rs}} \mathbb{I}_K u_{i''}^r$$

$$\frac{\partial \beta_{ir}^{SP_{ss}}}{\partial \mu_0^{RP_{rs}}} = 0 \qquad \frac{\partial \beta_{i''r}^{RP_{rs}}}{\partial \mu_0^{RP_{rs}}} = \lambda^{RP_{rs}} e_k \text{ if } \mu_{0,k}^{RP_{rs}} \neq 0$$

$$\frac{\partial \beta_{ir}^{SP_{ss}}}{\partial \lambda^{RP_{rs}}} = 0 \qquad \frac{\partial \beta_{i''r}^{RP_{rs}}}{\partial \lambda^{RP_{rs}}} = \beta_{i''r}^{RP_{rs}} / \lambda$$

where $e_k = (\underbrace{0, \ldots,}_{k-1} 1, \ldots, 0)'$. Therefore, substituting these into (21) will give us the jacobian functions for the $RP_{rs}$ and $SP_{ss}$ data.

Next, we can differentiate (21) further to obtain the Hessian matrix.

$$\begin{aligned}
\frac{d^2 LL}{d\Theta d\Theta'} &\simeq \sum_{i \in SP_{ss}} \sum_{r=1}^{R} \frac{d}{d\Theta'} \frac{l_{ir}}{\sum_{r=1}^{R} l_{ir}} \left( \sum_{t=1}^{T_i} (X_{ijt} - \sum_{k=1}^{J} l_{iktr} X_{ikt}) \right) \frac{d\beta_{ir}}{d\Theta} \\
&\simeq \sum_{i \in SP_{ss}} \sum_{r=1}^{R} \Bigg[ \frac{1}{\sum_{r=1}^{R} l_{ir}} \frac{dl_{ir}}{d\Theta'} \left( \sum_{t=1}^{T_i} (X_{ijt} - \sum_{k=1}^{J} l_{iktr} X_{ikt}) \right) \frac{d\beta_{ir}}{d\Theta} \\
&- \frac{l_{ir}}{\sum_{r=1}^{R} l_{ir}} \left( \sum_{t=1}^{T_i} \sum_{k=1}^{J} \frac{dl_{iktr}}{d\Theta'} X_{ikt} \right) \frac{d\beta_{ir}}{d\Theta} \\
&+ \frac{l_{ir}}{\sum_{r=1}^{R} l_{ir}} \left( \sum_{t=1}^{T_i} (X_{ijt} - \sum_{k=1}^{J} l_{iktr} X_{ikt}) \right) \frac{d^2 \beta_{ir}}{d\Theta d\Theta'}
\end{aligned}$$

(22)

$$\begin{aligned}
&- \frac{1}{(\sum_{r=1}^{R} l_{ir})^2} (\sum_{r=1}^{R} \frac{dl_{ir}}{d\Theta'}) \, l_{ir} \left( \sum_{t=1}^{T_i} (X_{ijt} - \sum_{k=1}^{J} l_{iktr} X_{ikt}) \right) \frac{d\beta_{ir}}{d\Theta} \Bigg] \\
&\simeq \sum_{i \in SP_{ss}} \Bigg( \sum_{r=1}^{R} \frac{l_{ir}}{\sum_{r=1}^{R} l_{ir}} \Bigg[ \frac{d\beta_{ir}}{d\Theta'} \left( \sum_{t=1}^{T_i} (X_{ijt} - \sum_{k=1}^{J} l_{iktr} X_{ikt}) \right)' \left( \sum_{t=1}^{T_i} (X_{ijt} - \sum_{k=1}^{J} l_{iktr} X_{ikt}) \right) \frac{d\beta_{ir}}{d\Theta} \\
&- \frac{d\beta_{ir}}{d\Theta'} \left( \sum_{t=1}^{T_i} \sum_{k=1}^{J} (X_{ikt} - \sum_{l=1}^{J} l_{iltr} X_{ilt})' l_{iktr} X_{ikt} \right) \frac{d\beta_{ir}}{d\Theta}
\end{aligned}$$

(23)

$$+ \left( \sum_{t=1}^{T_i} (X_{ijt} - \sum_{k=1}^{J} l_{iktr} X_{ikt}) \right) \frac{d^2 \beta_{ir}}{d\Theta d\Theta'} \Bigg] - \frac{dLL_i}{d\Theta'} \frac{dLL_i}{d\Theta} \Bigg)$$

Again, the format of the Hessian function is generic[11]. The $RP_{rs}$ and $SP_{ss}$ hessians differ due to the composition of their $X$ matrices, and the formulation of the $\frac{d^2\beta_{ir}}{d\Theta d\Theta'}$ matrix. For the $SP_{ss}$ data,

$$\frac{d^2\beta_{ir}}{d\Theta d\Theta'} = 0$$

For the $RP_{rs}$ data, $\frac{d^2\beta_{ir}}{d\Theta d\Theta'}$ is a $K\ X\ ((N_Z+1)K+5)\ X\ ((N_Z+1)K+5)$ cube such that

$$\begin{aligned}
\frac{d^2\beta_{ir}}{d\Theta d\Theta'} &= 0 \;\text{ if } \Theta, \Theta' \in \{\Delta_0, \Delta, \Sigma, \mu_0^{RP_{rs}}\} \\
&= 0 \;\text{ if } \Theta, \Theta' = \lambda^{RP_{rs}} \\
&= \frac{d^2\beta_i^{RP_{rs}}}{d\Theta d\lambda^{RP_{rs}}} \;\text{ if } \Theta \in \{\Delta_0, \Delta, \Sigma, \mu_0^{RP_{rs}}\}, \Theta' = \lambda^{RP_{rs}} \\
&= \frac{d^2\beta_i^{RP_{rs}}}{d\Theta d\lambda^{RP_{rs}}} \;\text{ if } \Theta = \lambda^{RP_{rs}}, \Theta' \in \{\Delta_0, \Delta, \Sigma, \mu_0^{RP_{rs}}\}
\end{aligned}$$

Substituting back into (23) gives us the full $((N_Z+1)K+5)\ X\ ((N_Z+1)K+5)$ Hessian matrix for the $RP_{rs}$ data. The Hessian matrix for the total log-likelihood is the sum of the Hessians for the $RP_{rs}$ and $SP_{ss}$ components.

[11] $l_{ijtr} = P_{ijt}(\Theta, u_i^r)$, i.e., likelihood of purchase for a generic product $j$ for individual $i$ at time $t$ with random draw $r$.

*Web Appendix B: Conjoint Data Details*

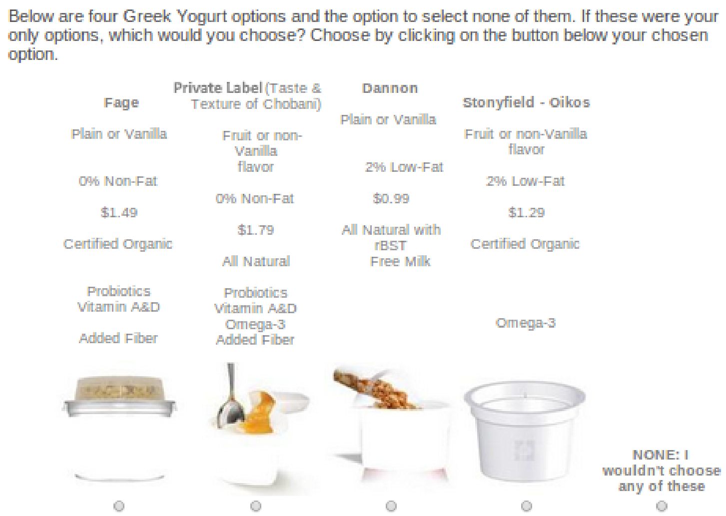


Figure 6: Sample Choice-based Conjoint Question

| Attributes | Marketplace<br>Levels | Conjoint<br>Levels | Attribute<br>Type |
|---|---|---|---|
| Cup-Style | Normal<br>Side-by-Side (SBS) Cup | Normal<br>Side-by-Side (SBS) Cup | Common |
| | | SBS Cup with Inclusions<br>Normal Cup with Inc. on Top | NTMA |
| Probiotic | | Yes/No | |
| Vitamins | | Yes/No | |
| Fiber | | Yes/No | |
| Omega-3 | | Yes/No | |
| Brand | | Private Label | |
| | Brown Cow[1] | | Actual |
| | Chobani[1]<br>Dannon<br>Fage[1]<br>Oikos[2] | Chobani<br>Dannon<br>Fage<br>Oikos | Common |
| 0% fat content | Yes/No | Yes/No | |
| Flavor | Plain/Fruit | Plain/Fruit | |
| Price | Continuous | $0.89, $0.99, $1.29, $1.49, $1.79<br>(Treated as continuous) | |
| Organic Content | Organic<br>All-Natural<br>rBst-Free | Organic<br>All-Natural<br>rBst-Free | Hidden |

[1] Brand coefficient represents Brand + All Natural
[2] Brand coefficient represents Brand + Organic

Table 14: Comparison of characteristics between Marketplace and Conjoint Survey

*Web Appendix C: Detailed Results*

| Attributes | $\Delta_0^{SP_{ss}}$ (S.E.) | | | $\Delta^{SP_{ss}}$ (S.E.) | | | $\sigma^{SP_{ss}}$ (S.E.) |
|---|---|---|---|---|---|---|---|
| | Constant | Log Weeks since last Greek yog. purchase | Avg. $ Sales in Greek yog. per store visit in past 90 days | Log No. store visits in Greek yog. in past 90 days | Inc. $\leq 75k$ | Family Size | Unobs. Het. SD |
| Inside Good | **2.11 (0.25)** | -0.77 (0.40) | **-0.20 (0.09)** | 0.18 (0.55) | 0.71 (0.50) | **1.37 (0.48)** | **3.18 (0.22)** |
| Brown Cow | | | | | | | |
| Chobani | **1.18 (0.13)** | **-0.46 (0.23)** | 0.00 (0.05) | -0.39 (0.26) | 0.11 (0.26) | **-0.64 (0.27)** | **1.56 (0.14)** |
| Fage | 0.40 (0.40) | -0.12 (0.63) | 0.02 (0.14) | -0.82 (0.82) | -0.54 (0.82) | 0.39 (0.77) | **2.76 (0.17)** |
| Oikos | **0.60 (0.19)** | -0.20 (0.35) | -0.05 (0.07) | 0.10 (0.40) | 0.01 (0.38) | -0.53 (0.37) | **1.58 (0.16)** |
| Private Label | **1.87 (0.36)** | -1.05 (0.56) | 0.14 (0.13) | -0.69 (0.73) | -0.12 (0.74) | 0.10 (0.70) | **1.88 (0.13)** |
| Zero-Fat | **1.10 (0.10)** | **0.44 (0.17)** | -0.04 (0.03) | 0.27 (0.20) | **0.52 (0.19)** | -0.11 (0.18) | **1.88 (0.10)** |
| Fruit | **0.31 (0.11)** | **-0.77 (0.18)** | 0.05 (0.04) | -0.33 (0.21) | 0.06 (0.22) | -0.33 (0.21) | **2.32 (0.11)** |
| Organic | 0.13 (0.12) | 0.25 (0.24) | 0.03 (0.05) | 0.07 (0.27) | -0.06 (0.25) | 0.20 (0.25) | **1.17 (0.12)** |
| All Natural | **-0.17 (0.07)** | **-0.29 (0.14)** | -0.02 (0.03) | -0.21 (0.16) | -0.16 (0.15) | 0.02 (0.15) | **0.39 (0.12)** |
| Probiotic | **0.12 (0.05)** | -0.13 (0.10) | 0.00 (0.02) | 0.01 (0.12) | 0.03 (0.11) | 0.03 (0.11) | 0.12 (0.09) |
| Vitamins | 0.07 (0.08) | 0.08 (0.15) | 0.02 (0.03) | -0.14 (0.17) | 0.04 (0.17) | -0.22 (0.16) | **0.44 (0.13)** |
| Omega-3 | -0.07 (0.08) | **-0.31 (0.15)** | 0.03 (0.03) | **-0.40 (0.18)** | 0.00 (0.17) | 0.10 (0.17) | **0.79 (0.10)** |
| Fiber | -0.03 (0.09) | -0.21 (0.15) | 0.01 (0.03) | -0.10 (0.19) | 0.22 (0.18) | -0.24 (0.17) | **0.86 (0.10)** |
| SBS Cup | **-1.37 (0.36)** | 0.67 (0.56) | 0.11 (0.13) | 0.02 (0.73) | 0.66 (0.74) | 0.08 (0.69) | **1.86 (0.16)** |
| SBS Cup Incl. | **-1.72 (0.12)** | -0.28 (0.20) | -0.03 (0.04) | **-0.68 (0.23)** | 0.29 (0.20) | 0.33 (0.21) | **1.84 (0.12)** |
| Normal Cup Incl. Top | **-1.43 (0.33)** | 0.66 (0.54) | -0.03 (0.12) | -0.15 (0.70) | 0.53 (0.72) | -0.02 (0.68) | **0.62 (0.21)** |
| Price | **-2.64 (0.20)** | 0.47 (0.35) | 0.08 (0.07) | -0.17 (0.42) | -0.57 (0.40) | -0.73 (0.39) | **1.25 (0.14)** |
| State Dependence | **2.52 (0.15)** | **-1.41 (0.26)** | **0.26 (0.05)** | **-0.68 (0.28)** | -0.47 (0.26) | **0.70 (0.26)** | **1.57 (0.16)** |
| No. of $SP_{ss}$ Choice occasions | | | | 6208 | | | |
| $SP_{ss}$ Log-Likelihood | | | | -6287.63 | | | |
| AIC | | | | 12827.26 | | | |
| BIC | | | | 13675.69 | | | |

Bold indicates significant at 5% level

Estimation done on $R = 5000$ simulated random normal draws.

Table 15: $SP_{ss}$ Only Estimation

| Attributes | $\Delta_0^{RP_{rs}}$ (S.E.) | | | $\Delta^{RP_{rs}}$ (S.E.) | | | $\sigma^{RP_{rs}}$ (S.E.) |
|---|---|---|---|---|---|---|---|
| | Constant | Log Weeks since last Greek yog. purchase | Avg. $ Sales in Greek yog. per store visit in past 90 days | Log No. store visits in Greek yog. in past 90 days | Inc. $\leq 75k$ | Family Size | Unobs. Het. SD |
| Inside Good | **-8.68 (0.19)** | **0.70 (0.29)** | **2.47 (0.29)** | -1.45 (0.85) | **0.79 (0.30)** | **1.60 (0.28)** | **6.31 (0.11)** |
| Brown Cow | **0.35 (0.14)** | 0.07 (0.19) | **0.29 (0.14)** | 0.61 (0.35) | -0.14 (0.24) | **1.64 (0.27)** | **2.01 (0.11)** |
| Chobani | **3.91 (0.12)** | **0.48 (0.19)** | 0.17 (0.17) | -0.10 (0.32) | **-0.63 (0.24)** | **-1.73 (0.26)** | **3.65 (0.09)** |
| Fage | **0.96 (0.33)** | 1.17 (0.61) | **2.18 (0.29)** | **-4.59 (0.87)** | **-4.28 (0.53)** | 0.85 (0.61) | **5.41 (0.28)** |
| Oikos | **-1.12 (0.29)** | **-2.20 (0.27)** | -0.03 (0.31) | **-2.50 (0.80)** | -0.10 (0.43) | **-0.91 (0.46)** | **6.43 (0.21)** |
| Private Label | | | | | | | |
| Zero-Fat | **0.72 (0.07)** | **-0.25 (0.12)** | 0.11 (0.12) | -0.11 (0.28) | -0.15 (0.14) | -0.15 (0.14) | **1.49 (0.06)** |
| Fruit | **1.94 (0.09)** | **0.90 (0.16)** | -0.19 (0.11) | **2.50 (0.46)** | **1.34 (0.16)** | **-0.96 (0.16)** | **3.79 (0.10)** |
| Organic | | | | | | | |
| All Natural | | | | | | | |
| Probiotic | | | | | | | |
| Vitamins | | | | | | | |
| Omega-3 | | | | | | | |
| Fiber | | | | | | | |
| SBS Cup | **-1.36 (0.42)** | **-2.36 (0.72)** | **-2.42 (0.24)** | -0.05 (1.49) | -0.15 (0.76) | 0.29 (0.66) | **4.02 (0.37)** |
| SBS Cup Incl. | | | | | | | |
| Normal Cup Incl. Top | | | | | | | |
| Price | -10.08 (0.20) | 0.24 (0.30) | -0.20 (0.29) | **7.14 (0.78)** | 0.30 (0.42) | **1.09 (0.40)** | **7.41 (0.14)** |
| State Dependence | **0.84 (0.06)** | **0.63 (0.09)** | **0.35 (0.05)** | **1.06 (0.19)** | -0.21 (0.14) | **0.32 (0.14)** | **1.64 (0.06)** |
| No. of $RP_{rs}$ Choice occasions | | | | 28731 | | | |
| $RP_{rs}$ Log-Likelihood | | | | -10743.07 | | | |
| AIC | | | | 21626.14 | | | |
| BIC | | | | 22204.74 | | | |

Bold indicates significant at 5% level

Estimation done on $R = 5000$ simulated random normal draws.

Table 16: $RP_{rs}$ Only Estimation

| Attributes | $\Delta_0^{SP_{ss}}$ (S.E.) | $\Delta^{SP_{ss}}$ (S.E.) | | | | | $\sigma^{SP_{ss}}$ (S.E.) |
|---|---|---|---|---|---|---|---|
| | Constant | Log Weeks since last Greek yog. purchase | Avg. \$ Sales in Greek yog. per store visit in past 90 days | Log No. store visits in Greek yog. in past 90 days | Inc. $\leq 75k$ | Family Size | Unobs. Het. SD |
| Inside Good | **-1.81 (0.09)** | **-0.26 (0.09)** | **-0.37 (0.05)** | 0.13 (0.17) | **-0.45 (0.14)** | **1.14 (0.16)** | **6.03 (0.20)** |
| Brown Cow | **-0.36 (0.10)** | 0.17 (0.13) | **0.09 (0.04)** | **0.68 (0.18)** | **0.35 (0.17)** | **0.39 (0.17)** | **1.29 (0.08)** |
| Chobani | **2.01 (0.08)** | 0.13 (0.11) | **0.10 (0.03)** | -0.12 (0.15) | **-0.74 (0.13)** | **-0.28 (0.13)** | **1.38 (0.05)** |
| Fage | **0.48 (0.14)** | 0.30 (0.20) | **0.11 (0.05)** | -0.31 (0.31) | -0.22 (0.24) | **0.47 (0.23)** | **2.40 (0.11)** |
| Oikos | **0.96 (0.13)** | -0.10 (0.22) | 0.03 (0.05) | -0.14 (0.29) | -0.29 (0.26) | -0.05 (0.26) | **0.62 (0.09)** |
| Private Label | **1.79 (0.19)** | -0.56 (0.33) | 0.11 (0.06) | -0.63 (0.41) | **0.93 (0.34)** | -0.65 (0.34) | **1.57 (0.13)** |
| Zero-Fat | **0.82 (0.04)** | **-0.16 (0.04)** | **-0.05 (0.01)** | **-0.19 (0.07)** | 0.06 (0.06) | **0.22 (0.06)** | **1.21 (0.05)** |
| Fruit | **0.67 (0.04)** | **0.16 (0.05)** | 0.01 (0.02) | **0.27 (0.08)** | 0.10 (0.07) | **0.16 (0.08)** | **1.65 (0.06)** |
| Organic | -0.08 (0.12) | 0.21 (0.20) | 0.07 (0.04) | -0.27 (0.26) | 0.14 (0.24) | 0.00 (0.24) | **1.62 (0.08)** |
| All Natural | -0.04 (0.06) | -0.12 (0.10) | **-0.05 (0.02)** | -0.11 (0.13) | **-0.48 (0.12)** | **0.33 (0.12)** | **0.40 (0.03)** |
| Probiotic | **0.18 (0.05)** | -0.10 (0.09) | 0.01 (0.02) | -0.01 (0.11) | 0.08 (0.10) | -0.02 (0.10) | 0.01 (0.08) |
| Vitamins | 0.08 (0.07) | 0.11 (0.13) | 0.04 (0.03) | -0.24 (0.15) | 0.15 (0.14) | **-0.33 (0.14)** | 0.24 (0.13) |
| Omega-3 | 0.01 (0.07) | **-0.37 (0.13)** | 0.04 (0.03) | **-0.60 (0.16)** | 0.15 (0.14) | -0.01 (0.14) | 0.22 (0.17) |
| Fiber | 0.05 (0.07) | -0.14 (0.13) | 0.03 (0.03) | -0.10 (0.15) | 0.20 (0.14) | -0.21 (0.14) | **0.36 (0.12)** |
| SBS Cup | **-1.30 (0.17)** | 0.01 (0.17) | -0.03 (0.05) | -0.25 (0.28) | **-1.12 (0.24)** | **1.01 (0.22)** | **1.60 (0.10)** |
| SBS Cup Incl. | **-1.47 (0.10)** | **-0.38 (0.18)** | 0.02 (0.03) | **-0.64 (0.21)** | 0.00 (0.20) | 0.38 (0.20) | **1.63 (0.13)** |
| Normal Cup Incl. Top | **-1.21 (0.17)** | 0.23 (0.25) | -0.07 (0.06) | -0.26 (0.34) | **-0.69 (0.30)** | **0.70 (0.28)** | **1.00 (0.13)** |
| Price | **-2.97 (0.12)** | -0.18 (0.13) | **-0.19 (0.05)** | **1.13 (0.23)** | **-1.35 (0.20)** | **0.52 (0.20)** | **3.34 (0.11)** |
| State Dependence | **0.97 (0.05)** | **0.12 (0.04)** | 0.03 (0.02) | **0.20 (0.06)** | -0.13 (0.07) | **0.71 (0.07)** | **1.30 (0.05)** |
| Scaling Factor $\lambda^{RP_{rs}}$ | | | | **1.80 (0.06)** | | | |
| No. of $SP_{ss}$ Choice occasions | | | | 6208 | | | |
| No. of $RP_{rs}$ Choice occasions | | | | 28731 | | | |
| $SP_{ss}$ Log-Likelihood | | | | -6774.42 | | | |
| $RP_{rs}$ Log-Likelihood | | | | -11046.74 | | | |
| Total Log-Likelihood | | | | -17821.15 | | | |
| AIC | | | | 35910.30 | | | |
| BIC | | | | 37044.12 | | | |

Bold indicates significant at 5% level

Estimation done on $R = 5000$ simulated random normal draws.

Table 17: $BENCH$ Estimation

| Attributes | $\Delta_0^{SP_{ss}}$ (S.E.) | $\Delta^{SP_{ss}}$ (S.E.) | | | | | $\sigma^{SP_{ss}}$ (S.E.) |
|---|---|---|---|---|---|---|---|
| | Constant | Log Weeks since last Greek yog. purchase | Avg. \$ Sales in Greek yog. per store visit in past 90 days | Log No. store visits in Greek yog. in past 90 days | Inc. $\leq 75k$ | Family Size | Unobs. Het. SD |
| Inside Good | **2.64 (0.22)** | -0.08 (0.39) | 0.08 (0.20) | 0.32 (0.33) | 0.12 (0.35) | **0.46 (0.19)** | **3.04 (0.08)** |
| Brown Cow | -0.31 (0.19) | **0.54 (0.17)** | -0.01 (0.12) | **1.00 (0.31)** | 0.36 (0.34) | 0.03 (0.18) | **1.10 (0.12)** |
| Chobani | **1.51 (0.10)** | **0.27 (0.12)** | -0.01 (0.05) | 0.01 (0.16) | -0.18 (0.31) | -0.33 (0.35) | **2.03 (0.06)** |
| Fage | **-0.43 (0.17)** | -0.45 (0.25) | **0.47 (0.14)** | **-1.60 (0.53)** | -1.22 (0.65) | 0.81 (0.50) | **2.83 (0.12)** |
| Oikos | **0.72 (0.13)** | -0.26 (0.22) | 0.12 (0.06) | -0.09 (0.29) | 0.04 (0.39) | -0.34 (0.47) | **1.51 (0.10)** |
| Private Label | **1.12 (0.26)** | **-1.06 (0.33)** | **0.27 (0.10)** | **-1.25 (0.49)** | -0.03 (0.80) | 0.37 (0.58) | **1.96 (0.13)** |
| Zero-Fat | **0.68 (0.05)** | **-0.23 (0.08)** | 0.00 (0.02) | -0.35 (0.22) | 0.07 (0.13) | 0.14 (0.24) | **1.16 (0.04)** |
| Fruit | **0.25 (0.11)** | 0.17 (0.29) | 0.04 (0.03) | 0.25 (0.34) | 0.13 (0.22) | **-0.27 (0.10)** | **1.63 (0.12)** |
| Organic | 0.11 (0.12) | 0.17 (0.21) | 0.06 (0.05) | -0.21 (0.25) | 0.15 (0.24) | 0.29 (0.30) | **1.22 (0.07)** |
| All Natural | -0.03 (0.06) | -0.23 (0.12) | -0.01 (0.03) | -0.20 (0.16) | -0.19 (0.13) | 0.01 (0.12) | **0.18 (0.02)** |
| Probiotic | **0.12 (0.05)** | -0.07 (0.10) | 0.01 (0.02) | -0.01 (0.12) | 0.06 (0.11) | 0.00 (0.11) | 0.04 (0.10) |
| Vitamins | 0.06 (0.07) | 0.08 (0.13) | 0.03 (0.03) | -0.28 (0.15) | 0.00 (0.15) | -0.29 (0.15) | 0.18 (0.18) |
| Omega-3 | -0.08 (0.08) | **-0.29 (0.14)** | 0.03 (0.03) | **-0.55(0.16)** | 0.07 (0.16) | 0.01 (0.16) | **0.60 (0.12)** |
| Fiber | 0.02 (0.08) | -0.17 (0.13) | 0.01 (0.04) | -0.08 (0.16) | 0.06 (0.15) | -0.16 (0.17) | **0.59 (0.15)** |
| SBS Cup | -0.66 (0.34) | **0.76 (0.22)** | **-0.22 (0.07)** | **0.94 (0.36)** | 0.08 (0.71) | -0.38 (0.59) | **1.69 (0.24)** |
| SBS Cup Incl. | **-1.65 (0.12)** | -0.23 (0.20) | 0.03 (0.04) | -0.46 (0.24) | 0.06 (0.21) | 0.34 (0.22) | **1.71 (0.11)** |
| Normal Cup Incl. Top | **-0.65 (0.20)** | **0.78 (0.25)** | **-0.25 (0.09)** | 0.67 (0.39) | 0.28 (0.74) | -0.35 (0.43) | 0.21 (0.39) |
| Price | **-2.26 (0.19)** | **-0.18 (0.09)** | -0.08 (0.15) | **1.24 (0.23)** | -0.16 (0.12) | -0.56 (0.50) | **2.62 (0.07)** |
| State Dependence | **2.20 (0.12)** | **0.24 (0.11)** | 0.12 (0.07) | **0.29 (0.13)** | **-0.18 (0.08)** | 0.11 (0.18) | **0.97 (0.05)** |
| Scaling Factor $\lambda^{RP_{rs}}$ | | | | **1.91 (0.08)** | | | |
| Additive Shifter $\mu_0^{RP_{rs}}$ for Inside Good | | | | **-7.20 (0.24)** | | | |
| Additive Shifter $\mu_0^{RP_{rs}}$ for Fruit | | | | **0.60 (0.11)** | | | |
| Additive Shifter $\mu_0^{RP_{rs}}$ for Price | | | | **-1.64 (0.33)** | | | |
| Additive Shifter $\mu_0^{RP_{rs}}$ for State Dep. | | | | **-1.88 (0.15)** | | | |
| No. of $SP_{ss}$ Choice occasions | | | | 6208 | | | |
| No. of $RP_{rs}$ Choice occasions | | | | 28731 | | | |
| $SP_{ss}$ Log-Likelihood | | | | -6486.15 | | | |
| $RP_{rs}$ Log-Likelihood | | | | -10769.83 | | | |
| Total Log-Likelihood | | | | -17255.98 | | | |
| AIC | | | | 34787.95 | | | |
| BIC | | | | 35955.62 | | | |

Bold indicates significant at 5% level
Estimation done on $R = 5000$ simulated random normal draws.

Table 18: $CPC$ Estimation

| Attributes | $\Delta_0^{SP_{ss}}$ (S.E.) | | | $\Delta^{SP_{ss}}$ (S.E.) | | | $\sigma^{SP_{ss}}$ (S.E.) |
|---|---|---|---|---|---|---|---|
| | Constant | Log Weeks since last Greek yog. purchase | Avg. \$ Sales in Greek yog. per store visit in past 90 days | Log No. store visits in Greek yog. in past 90 days | Inc. $\leq 75k$ | Family Size | Unobs. Het. SD |
| Inside Good | **2.35 (0.21)** | | | | **0.64 (0.08)** | -0.14 (0.07) | **2.19 (0.08)** |
| Brown Cow | 0.12 (0.06) | | | | -0.03 (0.13) | 0.05 (0.13) | **1.15 (0.05)** |
| Chobani | **1.81 (0.08)** | | | | 0.00 (0.11) | -0.01 (0.11) | **1.90 (0.07)** |
| Fage | **-0.81 (0.09)** | | | | **-0.94 (0.17)** | **-0.72 (0.16)** | **2.65 (0.10)** |
| Oikos | 0.33 (0.12) | | | | **-0.70 (0.23)** | -0.40 (0.23) | **1.36 (0.06)** |
| Private Label | **0.86 (0.14)** | | | | 0.18 (0.30) | 0.31 (0.29) | **1.73 (0.15)** |
| Zero-Fat | **0.81 (0.03)** | | | | **-0.18 (0.04)** | **-0.11 (0.03)** | **0.87 (0.03)** |
| Fruit | **0.34 (0.08)** | | | | **-0.12 (0.04)** | **-0.36 (0.04)** | **1.48 (0.05)** |
| Organic | 0.15 (0.11) | | | | 0.04 (0.22) | 0.25 (0.21) | **0.93 (0.05)** |
| All Natural | -0.09 (0.05) | | | | -0.18 (0.11) | -0.05 (0.11) | **0.07 (0.02)** |
| Probiotic | 0.09 (0.05) | | | | 0.04 (0.10) | 0.03 (0.10) | 0.14 (0.11) |
| Vitamins | 0.07 (0.07) | | | | -0.20 (0.14) | -0.22 (0.14) | **0.37 (0.10)** |
| Omega-3 | -0.07 (0.07) | | | | 0.08 (0.14) | -0.02 (0.14) | **0.37 (0.12)** |
| Fiber | 0.00 (0.07) | | | | -0.01 (0.14) | -0.18 (0.14) | **0.56 (0.12)** |
| SBS Cup | **-0.72 (0.07)** | | | | 0.09 (0.12) | **-0.63 (0.10)** | **1.74 (0.06)** |
| SBS Cup Incl. | **-1.57 (0.10)** | | | | 0.03 (0.21) | **0.52 (0.20)** | **1.68 (0.11)** |
| Normal Cup Incl. Top | **-0.53 (0.13)** | | | | 0.17 (0.24) | -0.09 (0.24) | **0.91 (0.17)** |
| Price | **-2.00 (0.19)** | | | | **1.00 (0.12)** | 0.20 (0.12) | **3.24 (0.11)** |
| State Dependence | **2.04 (0.12)** | | | | **0.10 (0.04)** | **0.16 (0.04)** | **0.76 (0.03)** |
| Scaling Factor $\lambda^{RP_{ss}}$ | | | | **1.74 (0.06)** | | | |
| Scaling Factor $\lambda^{RP_{rs}}$ | | | | **2.00 (0.08)** | | | |
| Additive Shifter $\mu_0^{RP_{ss}}$ for Inside Good | | | | **-5.01 (0.24)** | | | |
| Additive Shifter $\mu_0^{RP_{rs}}$ for Inside Good | | | | **-6.61 (0.28)** | | | |
| Additive Shifter $\mu_0^{RP_{ss}}$ for Fruit | | | | **0.56 (0.09)** | | | |
| Additive Shifter $\mu_0^{RP_{rs}}$ for Fruit | | | | **0.59 (0.09)** | | | |
| Additive Shifter $\mu_0^{RP_{ss}}$ for Price | | | | **-1.17 (0.09)** | | | |
| Additive Shifter $\mu_0^{RP_{rs}}$ for Price | | | | **-2.41 (0.23)** | | | |
| Additive Shifter $\mu_0^{RP_{ss}}$ for State Dep. | | | | **-1.30 (0.12)** | | | |
| Additive Shifter $\mu_0^{RP_{rs}}$ for State Dep. | | | | **-1.53 (0.12)** | | | |
| $SP_{ss}$ Log-Likelihood | | | | -6638.90 | | | |
| $RP_{ss}$ Log-Likelihood | | | | -31029.48 | | | |
| $RP_{rs}$ Log-Likelihood | | | | -11012.24 | | | |
| Total Log-Likelihood | | | | -48680.62 | | | |
| No. of $SP_{ss}$ Choice occasions | | | | 6208 | | | |
| No. of $RP_{ss}$ Choice occasions | | | | 29640 | | | |
| No. of $RP_{rs}$ Choice occasions | | | | 28731 | | | |

Bold indicates significant at 5% level

Estimation done on $R = 5000$ simulated random normal draws.

Table 19: $CPC$ Estimation with $RP_{rs}$, $RP_{ss}$ and $SP_{ss}$ data with Demographics Only

| Attributes | $\Delta_0^{SP_{ss}}$ (S.E.) | $\Delta^{SP_{ss}}$ (S.E.) | | | | | $\sigma^{SP_{ss}}$ (S.E.) |
|---|---|---|---|---|---|---|---|
| | Constant | Log Weeks since last Greek yog. purchase | Avg. $ Sales in Greek yog. per store visit in past 90 days | Log No. store visits in Greek yog. in past 90 days | Inc. $\leq 75k$ | Family Size | Unobs. Het. SD |
| Inside Good | **2.41 (0.21)** | **0.15 (0.05)** | **-0.22 (0.02)** | **0.75 (0.09)** | **0.60 (0.07)** | **0.17 (0.07)** | **2.35 (0.09)** |
| Brown Cow | 0.08 (0.06) | **0.31 (0.10)** | **0.07 (0.02)** | **0.63 (0.13)** | **0.46 (0.12)** | 0.02 (0.12) | **1.07 (0.04)** |
| Chobani | **1.48 (0.07)** | **-0.25 (0.10)** | **0.06 (0.02)** | **-0.44 (0.12)** | -0.05 (0.11) | **-0.67 (0.11)** | **1.78 (0.06)** |
| Fage | **-0.42 (0.10)** | **0.82 (0.13)** | **0.39 (0.03)** | 0.10 (0.17) | **-1.11 (0.16)** | -0.10 (0.15) | **2.62 (0.09)** |
| Oikos | 0.23 (0.12) | -0.24 (0.19) | -0.03 (0.05) | 0.32 (0.25) | 0.10 (0.24) | **-0.91 (0.23)** | **1.66 (0.07)** |
| Private Label | **1.05 (0.15)** | **0.68 (0.23)** | **0.22 (0.05)** | **0.64 (0.29)** | 0.02 (0.28) | 0.31 (0.27) | **1.68 (0.11)** |
| Zero-Fat | **0.77 (0.03)** | **0.05 (0.02)** | -0.01 (0.01) | **0.09 (0.03)** | **-0.19 (0.03)** | 0.05 (0.03) | **0.91 (0.03)** |
| Fruit | **0.29 (0.09)** | **0.09 (0.03)** | **0.03 (0.01)** | **-0.17 (0.04)** | **-0.07 (0.04)** | **-0.47 (0.04)** | **1.49 (0.05)** |
| Organic | 0.03 (0.11) | 0.19 (0.18) | 0.02 (0.04) | -0.27 (0.23) | 0.06 (0.22) | 0.08 (0.21) | **0.83 (0.04)** |
| All Natural | -0.05 (0.05) | -0.16 (0.09) | -0.03 (0.02) | -0.19 (0.12) | **-0.25 (0.11)** | -0.14 (0.11) | **0.27 (0.02)** |
| Probiotic | **0.10 (0.05)** | -0.08 (0.09) | 0.01 (0.02) | 0.03 (0.11) | 0.04 (0.10) | 0.05 (0.10) | 0.02 (0.10) |
| Vitamins | 0.06 (0.07) | 0.15 (0.12) | 0.02 (0.03) | -0.07 (0.15) | -0.09 (0.14) | -0.23 (0.14) | 0.30 (0.26) |
| Omega-3 | -0.07 (0.07) | **-0.30 (0.13)** | 0.02 (0.03) | **-0.45 (0.16)** | 0.03 (0.15) | -0.01 (0.15) | **0.57 (0.11)** |
| Fiber | 0.01 (0.07) | -0.18 (0.13) | -0.01 (0.03) | 0.01 (0.16) | -0.01 (0.15) | -0.12 (0.15) | **0.58 (0.14)** |
| SBS Cup | **-0.81 (0.09)** | **-0.87 (0.09)** | **-0.20 (0.02)** | **-0.64 (0.12)** | -0.12 (0.12) | **-0.56 (0.11)** | **1.88 (0.07)** |
| SBS Cup Incl. | **-1.62 (0.11)** | **-0.42 (0.20)** | 0.05 (0.04) | **-0.68 (0.22)** | 0.14 (0.21) | **0.57 (0.22)** | **1.71 (0.13)** |
| Normal Cup Incl. Top | **-0.70 (0.13)** | **-0.81 (0.20)** | **-0.20 (0.05)** | **-1.07 (0.25)** | 0.15 (0.24) | **-0.54 (0.24)** | **0.74 (0.15)** |
| Price | **-2.44 (0.20)** | **-0.56 (0.08)** | 0.04 (0.03) | **0.54 (0.14)** | **0.68 (0.11)** | 0.04 (0.11) | **3.11 (0.10)** |
| State Dependence | **2.17 (0.13)** | **0.09 (0.02)** | **0.03 (0.01)** | **0.08 (0.04)** | **0.15 (0.04)** | -0.04 (0.04) | **0.74 (0.03)** |
| Scaling Factor $\lambda^{RP_{ss}}$ | | | | **1.83 (0.07)** | | | |
| Scaling Factor $\lambda^{RP_{rs}}$ | | | | **2.00 (0.07)** | | | |
| Additive Shifter $\mu_0^{RP_{ss}}$ for Inside Good | | | | **-5.59 (0.27)** | | | |
| Additive Shifter $\mu_0^{RP_{rs}}$ for Inside Good | | | | **-6.77 (0.29)** | | | |
| Additive Shifter $\mu_0^{RP_{ss}}$ for Fruit | | | | **0.64 (0.09)** | | | |
| Additive Shifter $\mu_0^{RP_{rs}}$ for Fruit | | | | **0.51 (0.09)** | | | |
| Additive Shifter $\mu_0^{RP_{ss}}$ for Price | | | | **-1.05 (0.22)** | | | |
| Additive Shifter $\mu_0^{RP_{rs}}$ for Price | | | | **-1.30 (0.22)** | | | |
| Additive Shifter $\mu_0^{RP_{ss}}$ for State Dep. | | | | **-1.41 (0.13)** | | | |
| Additive Shifter $\mu_0^{RP_{rs}}$ for State Dep. | | | | **-1.84 (0.13)** | | | |
| No. of $SP_{ss}$ Choice occasions | | | | 6208 | | | |
| No. of $RP_{ss}$ Choice occasions | | | | 29640 | | | |
| No. of $RP_{rs}$ Choice occasions | | | | 28731 | | | |
| $SP_{ss}$ Log-Likelihood | | | | -6576.63 | | | |
| $RP_{ss}$ Log-Likelihood | | | | -30736.47 | | | |
| $RP_{rs}$ Log-Likelihood | | | | -10923.11 | | | |
| Total Log-Likelihood | | | | -48236.21 | | | |
| AIC | | | | 96758.41 | | | |
| BIC | | | | 98056.23 | | | |

Bold indicates significant at 5% level

Estimation done on $R = 5000$ simulated random normal draws.

Table 20: $CPC$ Estimation with $RP_{rs}$, $RP_{ss}$ and $SP_{ss}$ data with Demographics and RFM

*Web Appendix D: Maximum A-Posteriori Estimates For Managerial Decisions*

In this process of decision-making, we use the $CPC$ model coefficients to get Maximum a-Posteriori (MAP) estimates for each individual in the full random sample, which includes individuals that were not in the estimation sample. We refer to this sample as the $RF$ sample. These MAP estimates can only be obtained for the attributes observable in the actual purchases. We jointly maximize the information from individual-specific purchase decisions in the estimation period (refer Figure 2) and the estimated unobserved heterogeneity distribution. In that sense, we are integrating the prior (mixing) distribution and individual purchase histories to predict their future purchase probabilities (DeGroot 1970, Chapter 11). For the new to market and hidden attributes that can only be estimated from the conjoint data, we don't observe choices in this dataset to maximize the a-Posteriori likelihood. As a result, we rely on the observable heterogeneity (i.e., $Z_i$) to specify the parameters to individuals and integrate out the unobserved heterogeneity when computing the counterfactual simulations.

Given the model in equation (9) and a value of $\left(\beta_i^{RF}, \alpha_i^{RF}\right)$, which are a function of the $CPC$ parameters $\Theta$, $Z_i^{RF}$, and a draw $u_i$, the product purchase probability per visit for individual $i$ is

$$P_{ijt}^{RF}(\Theta, X^{RF}, p^{RF}) = \frac{\exp(X_{ijt}^{RF}\beta_i^{RF} + (p_{ijt}^{RF} - p_{i0t}^{RF})\alpha_i^{RF})}{1 + \sum_{k \in C_{it}^{RF}} \exp(X_{ikt}^{RF}\beta_i^{RF} + (p_{ikt}^{RF} - p_{i0t}^{RF})\alpha_i^{RF})},$$

where this probability is a function of the $CPC$ model parameters $\Theta = \{\Delta_0^{SP_{ss}}, \Delta^{SP_{ss}}, \Sigma^{SP_{ss}}, \lambda^{RP_{rs}}, \mu_0^{RP_{rs}}\}$, product attributes $X^{RF}$, and prices $p^{RF}$. The total data likelihood for individual $i$, given a draw $u_i$, is represented as

$$\mathcal{L}_i^{RF}(\Theta, X^{RF}, Z^{RF}, p^{RF}, u_i) = \prod_{t=1}^{T_i^{RF}} P_{i[j]t}^{RF}(\Theta, X^{RF}, Z^{RF}, p^{RF}, u_i) \tag{24}$$

where $[j]$ indicates the product chosen by individual $i$ in visit $t$. Given Equation (24) above, the MAP estimate maximizes the posterior likelihood of individual $i$ given by

$$
\begin{aligned}
u_i^{MAP}|\theta &= \arg\max_{u_i} \frac{\mathcal{L}_i^{RF}(\Theta, X^{RF}, Z^{RF}, p^{RF}, u_i) f(u_i)}{\int_{u_i} \mathcal{L}_i^{RF}(\Theta, X^{RF}, Z^{RF}, p^{RF}, u_i) f(u_i) du_i} \\
&= \arg\max_{u_i} \quad \mathcal{L}_i^{RF}(\Theta, X^{RF}, Z^{RF}, p^{RF}, u_i) f(u_i) \\
&\sim \arg\max_{u_i^r} \quad \mathcal{L}_i^{RF}(\Theta, X^{RF}, Z^{RF}, p^{RF}, u_i^r) f(u_i^r)
\end{aligned} \tag{25}
$$

Here, $f(\cdot)$ denotes the density of unobserved heterogeneity. Since the denominator is a constant for each individual, the maximum a-posteriori estimate is the one that maximizes the numerator, i.e., the combination of likelihood of purchase given the draw $u_i$ and the likelihood of drawing $u_i$ itself. In practice, we draw $R = 10000$ draws of $u_i^r$ for each individual, and pick the arg-max from among these.

To predict purchase probabilities for each individiual ($i$) for potential product offering $j$ and future purchase occasion ($t$), we use the MAP estimates $u_i^{MAP}$ and integrate over the distribution of preferences for the remaining $K_{NTMA} + K_{Hidden}$ new-to-market and hidden attributes, denoted below by simply the $NTMA$ subscript:

$$
\begin{aligned}
P_{ijt}(\Theta, X, Z, p, nu_i^{MAP}) &= \int\limits_{u_{i,NTMA}} P_{ijt}(\Theta, X, Z, p, u_i^{MAP}, u_{i,NTMA}) f(u_{i,NTMA}) \, du_{i,NTMA} \\
&= \int\limits_{u_{i,NTMA}} \frac{\exp(X_{ijt}\beta_i + (p_{ijt} - p_{i0t})\alpha_i)}{1 + \sum\limits_j \exp(X_{ijt}\beta_i + (p_{ijt} - p_{i0t})\alpha_i)} f(u_{i,NTMA}) \, du_{i,NTMA} \\
\text{where } (\beta_i, \alpha_i) &= Z\Delta + \Sigma\tilde{u}_i \\
\tilde{u}_i &= (u_i^{MAP}, u_{i,NTMA})'
\end{aligned} \tag{26}
$$

We simulate this predicted probability using $R = 10000$ draws of $u_{i,NTMA}$ from the standard normal distribution for every individual.

$$
P_{ijt}(\Theta, X, Z, p, u_i^{MAP}) \sim \frac{1}{R} \sum_{r=1}^{R} P_{ijt}(\Theta, X, Z, p, u_i^{MAP}, u_{i,NTMA}^r) \tag{27}
$$