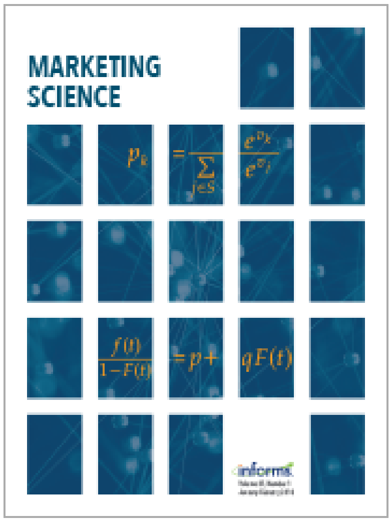


Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Monetizing Ratings Data for Product Research

Nino Hardt, Alex Varbanov, Greg M. Allenby

To cite this article:
Nino Hardt, Alex Varbanov, Greg M. Allenby (2016) Monetizing Ratings Data for Product Research. Marketing Science 35(5):713-726. https://doi.org/10.1287/mksc.2016.0980

MARKETING SCIENCE
Vol. 35, No. 5, September–October 2016, pp. 713–726
ISSN 0732-2399 (print) | ISSN 1526-548X (online)

http://dx.doi.org/10.1287/mksc.2016.0980

Monetizing Ratings Data for Product Research

Nino Hardt
Fisher College of Business, Ohio State University, Columbus, Ohio 43210, hardt.8@osu.edu

Alex Varbanov
The Procter & Gamble Company, Cincinnati, Ohio 45202, varbanov.ar@pg.com

Greg M. Allenby
Fisher College of Business, Ohio State University, Columbus, Ohio 43210, allenby.1@osu.edu

Features involving the taste, smell, touch, and sight of products, as well as attributes such as safety and confidence, are not easily measured in product research without respondents actually experiencing them. Moreover, product researchers often evaluate a large number of these attributes (e.g., >50) in applied studies, making standard valuation techniques such as conjoint analysis difficult to implement. Product researchers instead rely on ratings data to assess features for which the respondent has had actual experience. In this paper we develop a method of monetizing rating data to standardize product evaluations among respondents. The adjusted data are shown to increase the accuracy of purchase predictions by about 20% relative to existing methods of scale adjustment, leading to better inference in models using ratings data. We demonstrate our method using data from a large scale product use study by a packaged goods manufacturer.

Data, as supplemental material, are available at http://dx.doi.org/10.1287/mksc.2016.0980.

Keywords: prototype testing; experiential attributes; scale usage heterogeneity; Bayesian estimation; surveys
History: Received: July 7, 2014; accepted: October 5, 2015; Fred Feinberg served as the senior editor and Michel Wedel served as associate editor for this article. Published online in *Articles in Advance* May 3, 2016.

1. Introduction

Product developers are challenged by the need to prioritize and select the attributes that will most likely improve their product. It is common to have 50 to 100 attributes of interest to product researchers, each of which are qualitatively distinct in the sense that they require different engineering and manufacturing responses to achieve improvement. Many of these attributes need to be experienced by the consumer before they can be reliably evaluated. An example might be the desire to have a "clean feel" after applying an oral care product, or it leaving a pleasant "aftertaste." Attributes such as these are present in all forms of product research, ranging from services to durable goods.

The presence of experiential attributes in product research makes it difficult to use methods such as conjoint analysis (Green and Rao 1971) that rely on verbal descriptions of hypothetical products. Respondents in product research are typically exposed to a limited number of actual test products, often just one or two, and asked to provide evaluations with ratings scales after a trial period. Returning to oral care, attributes such as "mint" and "spearmint" can hardly cover the complexity of the taste-space that include flavor scents, intensity, and longevity. The goal of this research is to develop a reliable estimate of attribute value in situations where respondents are queried about a large number of attributes across a small number of test products.

Deciding how best to identify specific attributes for product improvement distinguishes product research from other marketing and psychology research areas that search for low-dimensional constructs and explanations of behavior. For product researchers, it is impossible to act based on a lower dimensional representation, as their goal is to achieve specific product experiences. In our study of oral care products, one may want to achieve a particular smooth feel of the teeth. This feeling is achieved via specific chemical formulations of the product, and it is not reasonable to reduce the high dimensionality of product research studies through the use of low dimensional statistical models (Luo et al. 2008) because of the product researcher's need to achieve a specific formulation.

Heterogeneity complicates analysis of scaled response data because respondents are known to use scales differently. Some respondents are yea-sayers and others are naysayers, some use the entire scale, and some just a fraction of it. Analysis of scaled response data requires some form of standardization of the responses so that they have common meaning before computing summary statistics and estimating model coefficients. An appropriate method of scale standardization for product research is the focus of this paper.

We develop a method of scale adjustment involving price to measure the dollar value a respondent ascribes to a scale point. Monetizing the rating scale facilitates a common and meaningful interpretation of ratings data. While it is not possible to measure the pain or the gain of one person relative to another, it is possible to translate either to an amount a person would be willing to pay for an improvement, and to use this information to give greater or less weight to respondents in calculating aggregate measures of performance and model coefficients. Our model involves collecting additional data on a subset of attributes to obtain a monetary value of a scale point, and demonstrating that the estimated conversion factor can be applied to all of the remaining attributes in an analysis.

The remainder of this paper is organized as follows. In Section 2 we review alternative approaches to scaling and discuss the strengths and weaknesses of existing methods. Our method of monetizing the scale is discussed in Section 3, and in Section 4 we demonstrate an empirical application involving a product test. Results are discussed in Section 5 and Section 6 provides concluding remarks.

2. Rescaling Ratings Data

The purpose of rescaling data is to improve its coherence among objects of measurement and respondents. Ratings data are particularly vulnerable to idiosyncratic meaning because of their ordinal nature. Data collected on a fixed-point rating scale provide relative information about the items under study and do not provide absolute information, such as the degree of belief in the statement associated with an item. It is difficult to relate data from heterogeneously scaled ratings unless some form of standardization takes place. We take the view that ratings data contain interval-scale information at the individual level, but that heterogeneity complicates its interpretation because the intervals are not common among respondents. Moreover, we only have limited individual-level information from respondents providing one or two sets of product evaluations. The challenge lies in standardizing rating scale evaluations to obtain meaningful "across people" inference on product attribute performance.

One approach to standardizing ratings data is to use information about the rated items themselves to provide the standardization. The cut-point model of Rossi et al. (2001), (RGA), for example, assumes a location (τ_i) and scale (σ_i) parameter affecting the proclivity of using certain portions of a fixed-point scale

$$x_{ij} = k, \quad \text{if } c_{k-1} \leq z_{ij} \leq c_k, \tag{1}$$

$$z_i \sim N(\mu + \tau_i \mathbf{1}, \sigma_i^2 \Sigma), \tag{2}$$

$$z_i^* = \frac{z_i - \tau_i \mathbf{1}}{\sigma_i} \sim N(\mu, \Sigma), \tag{3}$$

where the cut-points $\{c_k\}$ are fixed across respondents and provide a censoring mechanism for reconciling the discrete responses $x_{i,j}$ to a latent continuous variable $z_{i,j}$, where i indexes the respondents and j indexes the items. Subtracting τ_i and dividing by σ_i in Equation (3) puts the latent variable z_i^* on a scale that is common to all respondents. Variations on this approach allow for greater flexibility in the specification of the cut-points (Johnson 2003, Ying et al. 2006) and in specifying a mixture component model in Equation (3) that allows for haloing and straight-lining of responses (Büschken et al. 2013).

The cut-point model and other forms of statistical standardization assume that data are more meaningful when expressed on a statistical scale. Standardization replaces the original fixed-point scale with a measure of the number of standard deviations a rating is above or below a respondent's mean response. This approach controls for idiosyncratic interpretations of scale labels such as "very satisfied." However, statistical standardization does not address the issue that the original responses were low or high simply because the respondent disagreed or agreed with the propositions associated with the items. Alternatively, it may be the case that a respondent simply does not see any difference among the items for which they are queried, and truly feels they should be given the same rating. It is also not difficult to imagine instances where an opinion or experience was so exceptional that it truly deserved the highest mark on a scale. For these cases, a statistical approach to standardization is problematic because the location (τ_i) and scale (σ_i) adjustments lack meaningful content, i.e., the true origin and dispersion of the scale is lost. If τ_i, σ_i were truly capturing the scale use heterogeneity part of the observed data, their estimates must not only depend on the focal items given by the questionnaire design but on additional auxiliary items that help standardize the scaled responses.

We address the issue of conducting meaningful "across people" inference by monetizing the rating scale. If a respondent rates a product to be superior to another on a particular dimension, and is willing to pay for this difference, then they are given greater weight in our analysis relative to another respondent with the same ratings but lower monetary valuation. In the extreme case, where a respondent is not willing to pay for any change in performance of a product attribute, zero weight is assigned to their responses as a firm assesses the performance of their offering relative to others. Monetizing the rating scale allows for a direct interpretation of how much better one product performs than another on every item under study, regardless of the number of items in the survey.

The idea of rescaling data onto a "willingness-to-pay" (WTP) scale is not new. Sonnier et al. (2007) show that rescaling attribute importances in a multinomial logit model onto such a scale leads to improved fit and predictions, and Allenby et al. (2014) discuss the economic concept of WTP in conjoint analysis. Their approach addresses competitive reaction to the introduction of new attribute levels, which occurs further downstream in the product development process. Our use of monetization is intended to assist in the interpretation of ratings data; we do not claim that our scale is equivalent to economic WTP. Our money-scaled attribute evaluations, for example, are not guaranteed to sum up to a consumer's total WTP because the scale items may not be mutually independent. We therefore use the term "monetized scale" to avoid confusion with the economic concept of WTP.

One remedy for the lack of content to ground scale adjustment is to include additional items in the analysis that serve to anchor the responses. The classic response bias literature follows this approach, where constructs of response behaviors are developed and used to compute correction factors for the ratings (see, e.g., van Rosmalen et al. 2010, Baumgartner and Steenkamp 2001). Winkler et al. (1982), for example, collects responses to logically opposite items under the assumption that agreement with polar opposites indicates the presence of yea-saying. An adjustment parameter is thus estimated based on information external to the focal attributes.

Use of auxiliary data to identify a common scale origin is discussed by Böckenholt (2004) and applied to choice data by Bacon and Lenk (2012). Böckenholt (2004) considers paired comparisons, but his approach also applies to single-object evaluations. In both cases the scale origin is lost and can be recovered by auxiliary data. In some cases, the origin can be set by assumption. Bacon and Lenk (2012) consider choices made in a conjoint study, and use rating scale information to adjust the location of preferences.

Our approach to scale adjustment differs from Böckenholt (2004) and Bacon and Lenk (2012) in that we do not assume that people can provide absolute judgments in ratings data. We believe judgments reflected in ratings data are inherently relative; while use of auxiliary data helps ground the scale origin to an object external to the ratings data, that object itself is still relative. We propose standardizing the scale of ratings data, not its location or mean, so that movements along the scale for different items and different respondents have a common meaning. As shown in Section 3, we do this by introducing auxiliary data that relates the scales to an objective common measure; we do not address an absolute anchor for problems involving comparisons.

3. Monetized Scale Adjustment

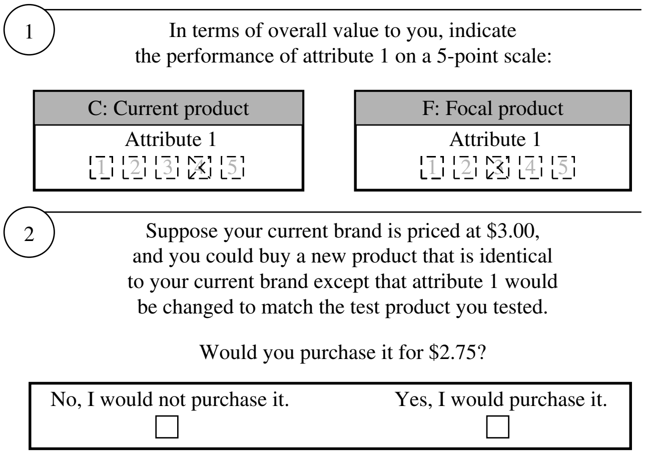
Our goal is to measure the worth of a scale point change in dollars. Conversion to a monetary metric is particularly useful in product analysis because it is aligned with the task of profit maximization. Note however, that a monetary scale is not useful with objects that are not for sale, such as when ratings are used to assess attitudes and opinions of one's self (e.g., I am a spend-thrifty shopper), or in the evaluation of nonprofit pursuits.

Our model requires value-scaled ratings (i.e., respondents are asked to rate attributes "in terms of overall value" to them) for the focal product (F) and two additional pieces of auxiliary information, i.e., (i) value-scaled ratings for a competing (C) product for a subset of the attributes (of length j); and (ii) binary choices between the competing product and a hypothetical product that integrates one feature of the focal product into the competing product.

We use the auxiliary data to estimate the relationship between value-scaled ratings and price that is useful for rescaling the focal product evaluations. In our empirical application, we find that an auxiliary data set comprising additional ratings and choices provides reliable estimates of the monetary value-scaled relationship. Figure 1 illustrates an example question with ratings and choice task.

Note that our method does not establish a product's worth in terms of its location on the scale. It is often unrealistic to expect that respondents can state the dollar value, or reservation price, of a product in an absolute sense. For example, the questions, "What would you pay for a digital point-and-shoot camera?" and "What would you pay for nursing home care?" are difficult to answer without information about market prices, regardless of the amount of product information provided. We believe it is more reasonable to expect respondents to know if a particular improvement of a product attribute is worth a specified amount of money. Our model thus yields a scalar adjustment parameter β_i, which re-scales the difference between evaluations of F and C to take on monetary meaning.

Figure 1 Questionnaire



1 In terms of overall value to you, indicate the performance of attribute 1 on a 5-point scale:

C: Current product	F: Focal product
Attribute 1	Attribute 1
1 2 3 ☒ 5	1 2 ☒ 4 5

2 Suppose your current brand is priced at $3.00, and you could buy a new product that is identical to your current brand except that attribute 1 would be changed to match the test product you tested.

Would you purchase it for $2.75?

No, I would not purchase it. ☐ Yes, I would purchase it. ☐

In addition, we do not attempt to force respondents to disentangle importance and performance for our calibration questions. On the contrary, we propose using value scaling to encourage respondents to express the difference in their ratings in terms of a dollar value. Value can be understood as a combination of importance and performance; we simply seek to monetize changes in scale ratings for attributes. We use side-by-side evaluations of a focal and a competing brand, asking for the value each attribute adds to the overall experience. Repeated observations of rating scale differences, and associated choices, provide the data needed to estimate an individual-level scale adjustment coefficient.

Formal Model

The choice probability for the new product with one feature of the focal product added (F) over the competitive product (C) is expressed as

$$\begin{aligned}\Pr(\mathrm{F}) &= \Pr(x_{\mathrm{F}}\beta - \ln p_{\mathrm{F}} + \varepsilon_{\mathrm{F}} > x_{\mathrm{C}}\beta - \ln p_{\mathrm{C}} + \varepsilon_{\mathrm{C}})\\ &= \frac{\exp[(x_{\mathrm{F}}\beta - \ln p_{\mathrm{F}})/\sigma]}{\exp[(x_{\mathrm{F}}\beta - \ln p_{\mathrm{F}})/\sigma] + \exp[(x_{\mathrm{C}}\beta - \ln p_{\mathrm{C}})/\sigma]}\\ &= \frac{\exp[(x_{\mathrm{F}}\beta - \ln p_{\mathrm{F}})/\sigma]}{\exp[(x_{\mathrm{F}}\beta - \ln p_{\mathrm{F}})/\sigma] + \exp[(x_{\mathrm{C}}\beta - \ln kp_{\mathrm{F}})/\sigma]}\\ &= \frac{\exp[((x_{\mathrm{F}} - x_{\mathrm{C}})\beta - \ln k)/\sigma]}{1 + \exp[((x_{\mathrm{F}} - x_{\mathrm{C}})\beta - \ln k)/\sigma]},\end{aligned} \tag{4}$$

where x_{F} and x_{C} are the scale ratings for the focal (F) and competing (C) offers, and k is the price discount/premium expressed as a percentage of the cost of the new product. We assume Extreme Value distributed errors ε_{F} and ε_{C} with mode zero and scale σ. In this expression, β is the monetary value, as a percentage of the reference price, of a one unit rating scale difference. The parameter σ is an error term parameter that allows us to deal with inconsistent choices. The likelihood is similar to the one used by Sonnier et al. (2007), where the price coefficient is fixed at 1, and σ is estimated. However, we use the natural logarithm of price. Because all x are value-scaled, i.e., rated in terms of overall value to the respondent, β should not take on a negative value. The relationship between the price coefficient and error scale in discrete choice models is discussed by Chandukala et al. (2008).

Each respondent makes choices across a number of products that differ on one attribute and price, allowing estimation of individual-level monetary value (β) and error scale (σ) parameters using Bayesian methods. Across individuals, we assume that parameters are distributed log-normal (Equation (5)) to insure positivity of the estimates. The assumption of a log-normal distribution is also advantageous in that the prior is invariant as to our likelihood based on Sonnier et al. (2007), or a traditional likelihood that introduces a separate price coefficient instead of estimating the error scale σ. If products are rated the same on an attribute, there is only information about the error term. As discussed by Meade and Craig (2012), Krosnick (1991), the presence of an error term is important in addressing careless and otherwise effort-minimizing respondents

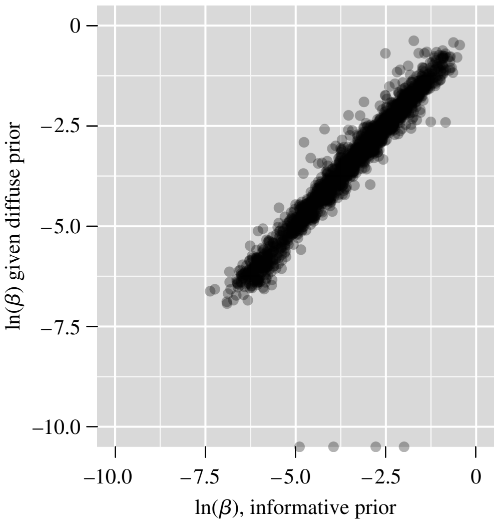
$$\begin{pmatrix}\ln\beta_i\\ \ln\sigma_i\end{pmatrix} \sim \mathrm{MVN}(\mu, \Sigma). \tag{5}$$

Estimation is performed using a Markov Chain Monte Carlo (MCMC) algorithm. The individual-level parameters are updated using the Metropolis–Hastings algorithm (Metropolis et al. 1953, Hastings 1970); the bivariate distribution of individual $\{\beta_i\}$, $\{\sigma_i\}$ is updated using a Gibbs sampler (Geman and Geman 1984).

The logarithmic specification of $\{\beta_i\}$, $\{\sigma_i\}$ has to be kept in mind when specifying the prior distributions. Consider, for example, the prior $\mu \sim N(-3, 1/3)$. The expected values were chosen to be close to zero $\exp(-3) \approx 0.05$. Here, 0.05 is equivalent to a 15 cent value of a scale point difference in our application below. The variance of the prior distribution corresponds to a 95% interval of $(0.02, 0.12)$ which is equivalent to a 6 to 39 cent value. Given the distribution of differences perceived, the number of attributes (21) and the base price of \$3, we assess this to be a reasonable range, leaving room for different weights of the responses. Using a less informative prior $\mu_i \sim N(-3, 10)$ has little impact on the estimated parameters as shown in Figure 2. However, we find it desirable to shrink extreme β_i more towards reasonable values with the more informative prior.

Finally, the adjustment procedure is straightforward: If both products F and C have been evaluated in terms of an attribute, the rescaled value of the improvement

Figure 2 **Estimates of β_i Given Two Different Prior Distributions**



is $(x_\mathrm{F} - x_\mathrm{C})\beta$. For attributes without dual evaluations, we can approximate the adjusted score by computing $(x_\mathrm{F} - \sum_k x_{k\mathrm{C}}/k)\beta$, where j is the size of the subset of attributes used for calibration.

Benchmark Models

We consider three alternative models in our empirical application below, in which we propose use of a weighted difference, $(x_\mathrm{F} - x_\mathrm{C})\beta$, in scale ratings as a basis for interpreting ratings data in product analysis. A common industry practice is to simply use the scale response alone, (x_F), which assumes that respondents can correctly and consistently use the scales provided in the questionnaire. Alternatively, analysis can proceed using the raw differences in scale responses $(x_\mathrm{F} - x_\mathrm{C})$, which acknowledges that differences in the ratings are important for analysis. Use of raw differences is consistent with the view that ratings are inherently relative and that inference should be based on improvements relative to an alternative product. Our uses of β to weight the difference in ratings takes this perspective a step further by giving greater weight to differences associated with greater monetary gain. We believe such weighting is appropriate in for-profit pursuits in which alternatives can be purchased in a marketplace.

A third alternative model for comparison is the model of scale use heterogeneity by Rossi et al. (2001). This model derives its adjustment to the scale based on the same items used in the evaluation. It does not use auxiliary data, nor does it advocate using differences in response. The scale adjustment is based on the idea that the relevant information for product analysis is the relative importance of the rated product attributes, where inference is made relative to the other attributes. We believe that such an adjustment is useful for identifying the attributes that are more important to a respondent, but since the adjusted rating is not linked to an object or behavior external to the product, its use in guiding management to improve an offering is limited. Internal scaling does not identify whether some aspect of a product is worth improving. It only identifies which aspect of a product is most important relative to the other items.

4. Empirical Application

We examine monetary scale adjustment performance using data from a product research study of a major packaged goods manufacturer. The study involves home testing of variants of an oral care product over a two-week period, and then answering a series of questions about their impression of the new product. Examples include consumer impressions of attributes such as: taste/flavor, foam, aftertaste, freshen breath, feeling clean, enjoyable experience, ease of use, cavity prevention, and trust. A follow-up questionnaire was administered to respondents a few days after the original survey to obtain information about the respondent's current product and additional information needed for our model. Unfortunately we cannot reveal specific items in our analysis below because of confidentiality requirements.

Thirteen variants of the product were investigated. Each respondent was randomly assigned to receive one of the test products for evaluation. Therefore in each of the 13 legs of the study, one focal product (the provided test product) and various competing products (the ones previously used by the respondents) were evaluated. While the main study included more than 50 different product attributes, the follow-up study included dual ratings (focal test product and previously used product) on 21 attributes. Binary choice responses were used to obtain estimates of β and σ for each respondent. In addition, the follow-up questionnaire included a validation question that involved choosing between the focal and competing product at five different price levels. The advantage of this choice-based WTP question is that it is unaffected by scale use heterogeneity and can thus serve as a benchmark. Throughout, we set prices to mimic the marketplace to make the survey reasonable.

A total of 3,236 consumers completed the follow-up survey, and 2,237 respondents provided usable data for analysis.[1] Respondents were excluded from the study if they did not provide answers to all questions in both the original and auxiliary surveys, straight-lined their answers or believed that the focal product and competitive product performed identically on all 21 scale items. While the full set of items and the exact wording must remain confidential, we can give a few examples: taste/flavor; freshen breath; feeling clean; foam; enjoyable experience; easy to use; trust; and best for health of mouth. For the 21 items, we specifically asked respondents to evaluate these items as to the value to them (see Figure 1).

Distribution of the number of different ratings for the focal (F) and competitive (C) product across the 21 items for each respondent is shown in Figure 3. Approximately 100 customers evaluated both brands equally on all items, however, they would make different decisions on the auxiliary choice task. This allows us to estimate only the error term for these respondents. However, most customers see a number of differences, providing enough information to estimate two individual-level parameters.

There are notable variations between the items in terms of how often respondents saw differences

[1] Respondents had to access the auxiliary study separately online, since the manufacturer could not include it as part of the main questionnaire. This is the primary reason for the decline in respondents.

Figure 3 Number of Different Ratings

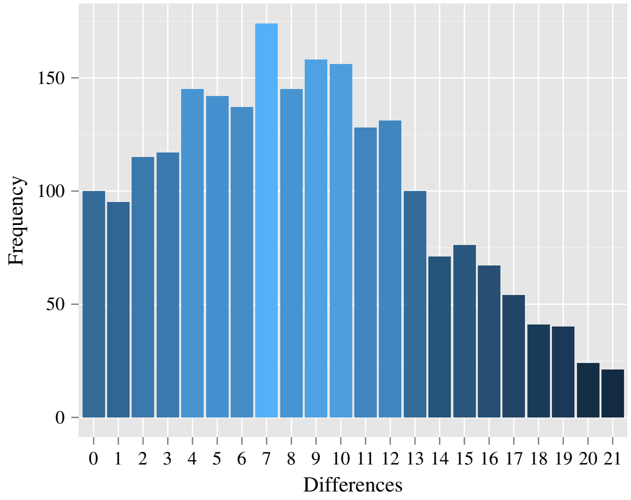
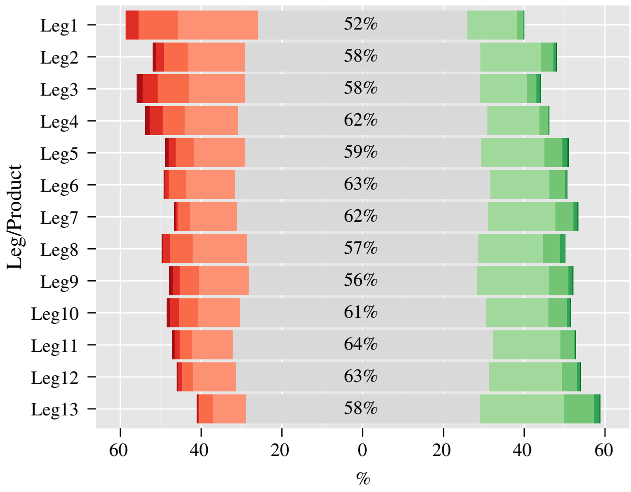
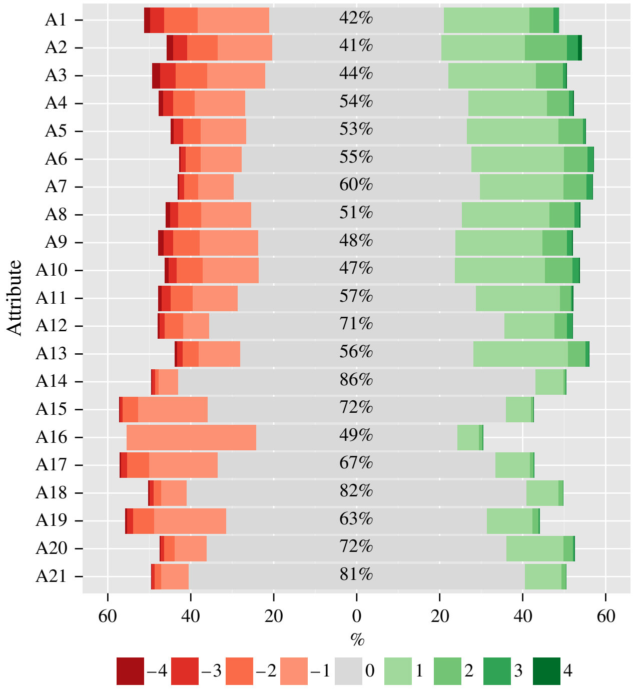


Figure 5 Scale Point Differences by Groups



between the two products. This suggests that an intelligent choice of items can actually increase the efficiency of our approach. In Figure 4 we see that some attributes led 60% of the respondents to provide different responses for the two products (Attributes 1 and 2), while approximately 80% thought that Attribute 14 was the same for both the focal and competitive product. These particular findings are not surprising; attribute 14 is based on a very generic statement, while attributes 1 and 2 focus on clearly distinguishable product characteristics. Attribute 16 is "trust in the product," and consequently many people rate their familiar current product higher than the unknown test product. Our findings suggest that factual, easily distinguishable attributes should be used to keep the length of the auxiliary questionnaire small.

Figure 4 Scale Point Differences by Attribute



Notes. For each attribute, the percentage indicates the proportion of respondents who rate the focal and competitive products equally. The bars in green indicate superior evaluations of the focal product, and the bars in red indicate the opposite.

The shares of no-difference versus difference seen appear to be equally distributed across the 13 legs of the study (Figure 5). However, some test products triggered more positive differences, while others lead to more negative comparisons. Independent of the actual test product performance, we always obtain enough information to estimate the scale point value and error variance in our model.

As expected, a less favorable side-by-side evaluation can be compensated by a price discount. The effect of price and rating scale difference on the share of choices in favor of the test product is shown in Figure 6. Given an equal rating on an attribute and given an equal price, the test product is chosen 59% of the time (see center bar in the graph). A slight increase in price by 15 cents almost halves the share of choices in favor of the test product's attribute. Given a 1 scale point improvement, however, again 59% of the time respondents chose the test product's feature at the higher price ($3.15). These results are in line with our assumption: Respondents react to price changes in a choice task, revealing the true strength of their preferences.

Figure 7 displays the distribution of average ratings across the 21 attributes, broken down by the respondents' WTP. The averages are computed across the focal and competing products, capturing the tendency to yea-saying versus naysaying. If there were a relationship between yea-saying and "over-committing" to large WTP, then the box plot would reveal that relationship. However, it is apparent that there is no systematic relationship between WTP and rating average.

Figure 6 **Choices of Test Product Given Price and Rating Difference**

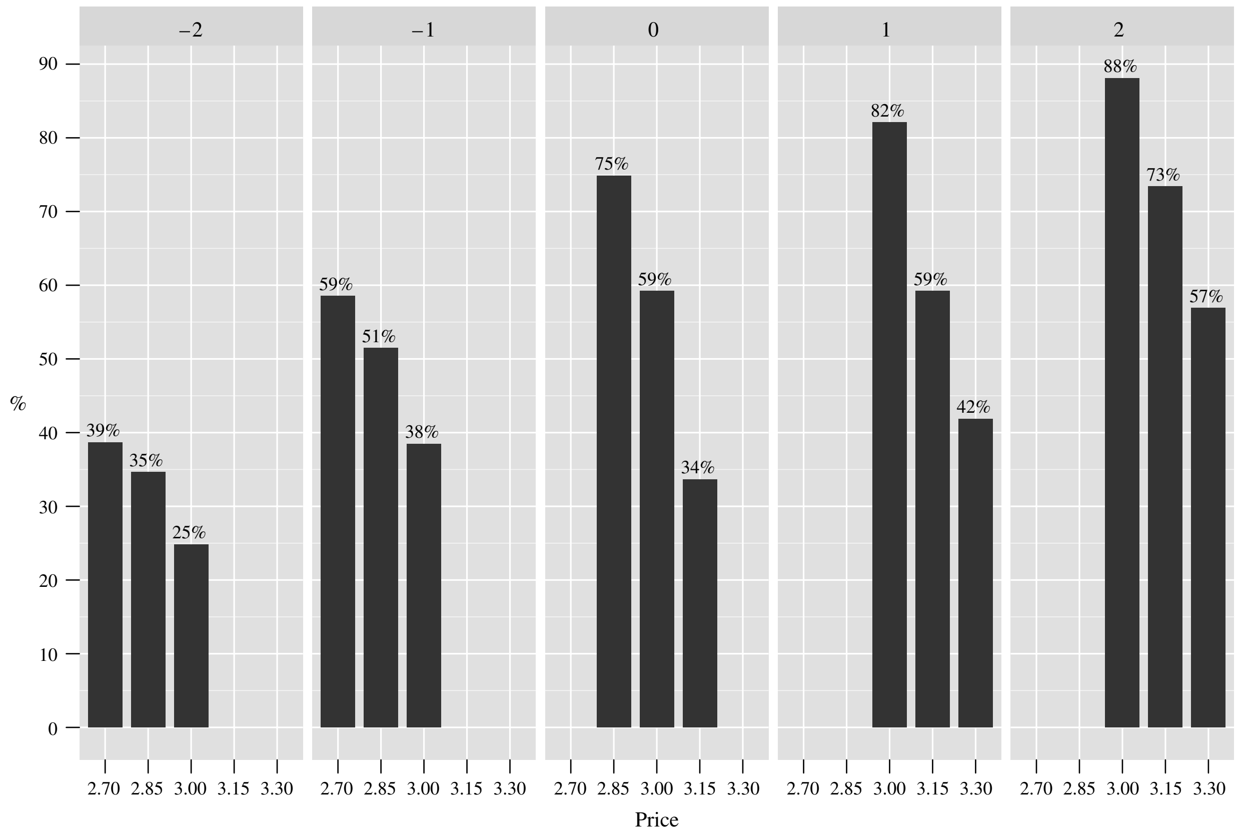
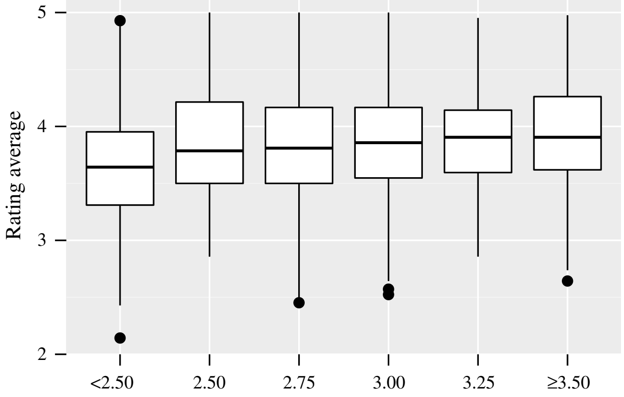


Figure 7 **Average Rating (Across 2 × 21 Evaluations) by WTP (Focal Product)**



5. Estimates

Posterior estimates of model parameters are based on 60,000 iterations of the Markov chain, with 30,000 draws used to burn-in the chain. The remaining 30,000 draws were used to conduct inference. Summaries of the model parameters are presented in Tables 1 and 2. We find that the Markov chain converges quickly and to the same posterior distribution from multiple starting points.

Table 1 **Estimates of 1st Stage Prior Parameters μ, Σ**

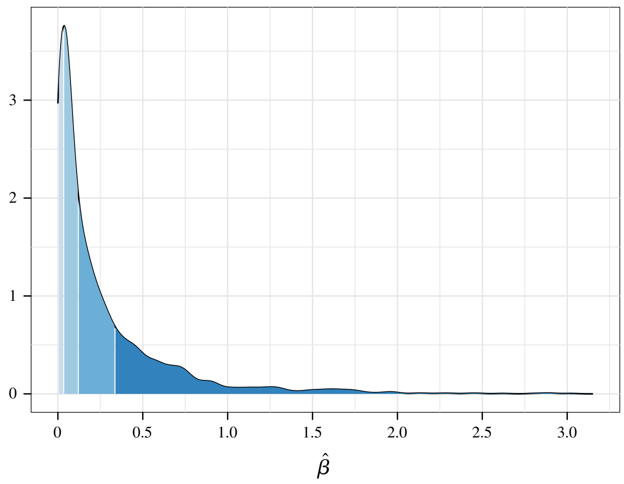
Parameter	Mean	Mode	2.5%	97.5%
μ_1	−2.80	−2.77	−3.32	−2.67
μ_2	−2.49	−2.50	−2.58	−2.33
Σ_{11}	3.12	2.86	2.49	7.71
Σ_{12}	1.97	1.90	1.65	2.77
Σ_{22}	1.88	1.80	1.56	2.77

Table 2 summarizes the individual-level estimates of $\ln \beta_i$ and $\ln \sigma_i$. We find that the individual-level estimates closely match corresponding estimates of the hyper-parameters reported in Table 1, indicating that the log-normal distribution of heterogeneity is correctly specified (Allenby and Rossi 1998). The posterior summaries of the individual-level estimates are

Table 2 **Summaries of $\ln(\beta_i)$, $\ln(\sigma_i)$ Posterior Distributions**

Posterior	Parameter	Mean	Min	Max	2.5%	97.5%
Mean	$\ln \beta_i$	−2.80	−6.35	0.43	−5.52	−0.33
	$\ln \sigma_i$	−2.49	−5.06	−0.68	−4.71	−1.03
Median	$\ln \beta_i$	−2.81	−6.22	0.31	−5.38	−0.44
	$\ln \sigma_i$	−2.58	−4.99	−0.73	−4.61	−1.12
Mode	$\ln \beta_i$	−2.86	−6.02	0.35	−5.20	−0.63
	$\ln \sigma_i$	−2.75	−4.96	−0.75	−4.50	−1.24

Figure 8 (Color online) Distribution of Posterior Means of β



invariant to whether the posterior mean, median or mode is used as a summary.

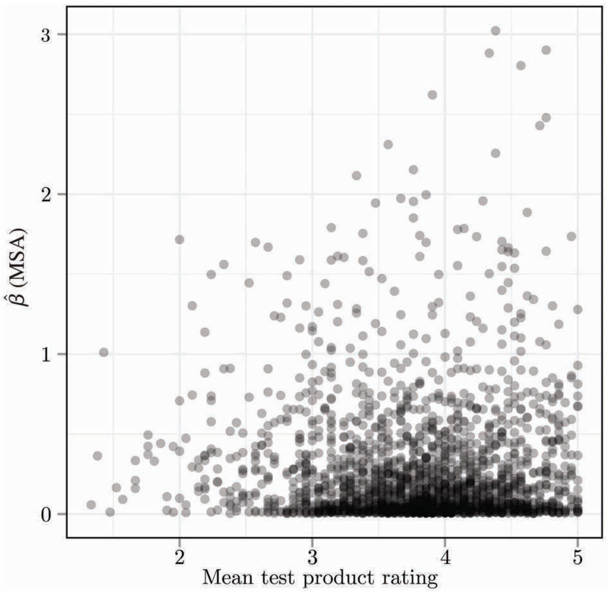
Figure 8 displays the distribution of respondents' posterior means of β. The shading in the figure corresponds to the quartiles of the distribution. There is a substantial amount of heterogeneity, ranging from respondents with close-to-zero value of scale point differences to respondents with estimates of β greater than 0.40 (80% percentile). Because price in the model is measured on a logarithmic scale, the coefficient estimate of 0.40 is interpreted to mean that a unit scale change is equal to a 40% change in the price of the good. We use a base price of \$3.00 in our analysis. Therefore, a coefficient of this magnitude indicates a scale worth of $\$3.00 \times 0.40 = \$1.20$. Figure 8 indicates that few respondents have such extreme values. The majority of respondents have estimates of less than 0.05, which corresponds to a scale point valued at \$0.15 or less. Respondents with a large β estimate tend to use the scale more conservatively, so that even a slight change on the rating scale is equivalent to a large change on the monetary scale.

To better understand the scale adjustment parameter β, we plot it against the mean evaluation of the test product in Figure 9. We do not expect to see a significant relationship between the two, as the heterogeneity of the dollar-value scale link should be independent of the absolute evaluation of a product. The fact that some respondents receive a more appealing test product should not influence their use of the value scale. Moreover, the tendency to respond high or low on a rating scale should also be independent of someone's WTP. Our data confirm this hypothesis.

Validation

Validating ratings data analysis is challenging, as only the observation of a "true" buying propensity or actual behavior can validate the model. By contrast, correlation with overall measures such as "overall satisfaction" can be affected by factors such as scale-use bias, as the dependent variable may itself be subject to the same scale use tendencies affecting the other scale items. Adjusting both the dependent and independent variables, however, is not a fair comparison of models.

Figure 9 $\{\beta_i\}$ vs. μ_{test}



We therefore asked respondents whether they would buy the test product at different price levels. The highest acceptable price level was taken to be their WTP, which serves as an indicator of the true buying intention. Respondents were presented with five different price levels, given that their current product was fixed at \$3. This gives rise to six different WTP categories. Note that the validation task is not affected by scale use heterogeneity and can thus be used to compare any scale adjustment model.

Table 3 shows the distribution of WTP levels among the respondents. A total of 299 respondents (13%) are in the top-box, stating a WTP of \$3.50 or higher. Also collected as part of the survey was a traditional purchase intention measure (I definitely would not buy it (1)—I definitely would buy it (5)). Table 4 displays a comparison of these two measures; clearly, there is heterogeneity in the WTP measure in each of the purchase intent (PI) responses. This heterogeneity indicates that price plays an important role in purchase

Table 3 Validation

Price	Cases	(%)	Cumulative (%)
<2.50	505	23	23
2.50	141	6	29
2.75	398	18	47
3.00	611	27	74
3.25	283	13	87
≥3.50	299	13	100

Table 4 WTP-Task vs. Purchase Intent (PI) Measure

	WTP						
PI	< $2.50	$2.50	$2.75	$3.00	$3.25	≥ $3.50	Sum
1	75	3	0	2	1	0	81
2	170	16	32	8	4	4	234
3	170	64	159	129	29	19	570
4	83	46	175	351	157	121	933
5	7	12	32	121	92	155	419
Sum	505	141	398	611	283	299	2,237

Note. PI: I definitely would not buy it (1)—I definitely would buy it (5).

decisions even when respondents are enthusiastic about purchasing the product and give a product a score of five on the purchase intention scale.

We use WTP estimates to validate the proposed and alternative benchmark models. Figure 10 displays the predictive performance of the alternative models in terms of a hit-rate of correct classification of respondents assigned to each WTP category. The hit rates were calculated as the fraction of correct classification of respondents into each category when sorted by the sum of the 21 items for the focal product.

For example, the "Raw" score hit-rate was calculated by first scoring respondents based on the sum of their raw evaluations of the focal product ($W_{\text{raw}} = \sum_k x_{\text{F}_k}$). The top 13% of respondents were assigned to the highest group, the next 13% were assigned to the second highest group, and so on in accordance with the fraction of respondents reported in Table 3. Sorting for the "Differences" model is based on the summed difference of the focal minus the competitive product ($W_{\text{diff}} = \sum_k (x_{\text{F}_k} - x_{\text{C}_k})$); sorting for the "MSA" model weighted the difference by the value of β ($W_{\text{MSA}} = \sum_k \hat{\beta}(x_{\text{F}_k} - x_{\text{C}_k})$). For the RGA model, which Rossi et al. (2001) use for external validation of a satisfaction study, sorting is based on the latent evaluations (z), ($W_{\text{RGA}} = \sum_k \hat{z}_{\text{F}_k}$). The hit rate was then calculated as the fraction of respondents in each group who also expressed the WTP of the group. A perfect scale adjustment model would find the 299 respondents with the highest rating of the focal product to be those who would spend $3.50 or more for it. The MSA model yields the highest hit rate, except in the case of respondents with a WTP of $\geq$ $2.50 and $<$ $2.75. Overall, the improvement is dramatic, yielding a 18% higher hit-rate.

Figure 10 (Color online) Validation (21 Items)

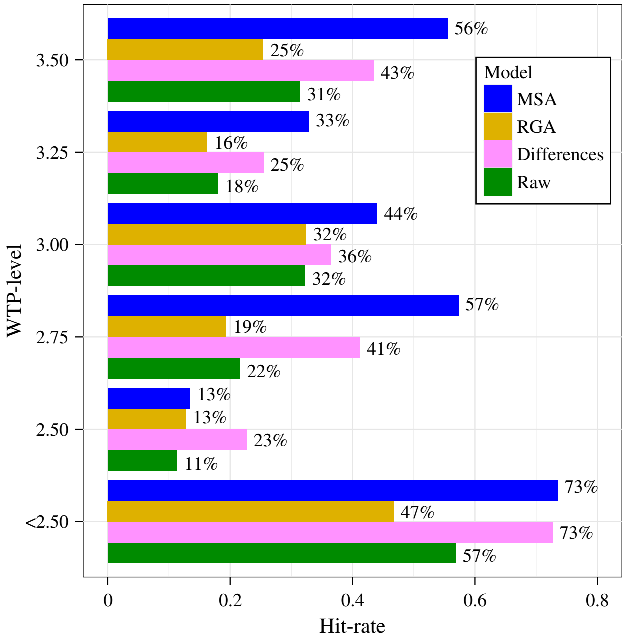
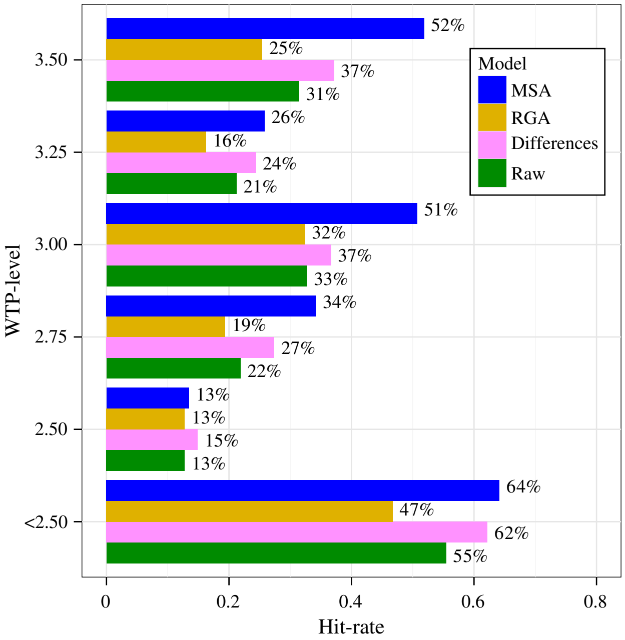


Figure 11 (Color online) Validation (57 Items)



Having identified the scale point value for each respondent, we can also adjust additional items from the study without the need for side-by-side evaluations. There were 57 additional questions in our study that have been asked during the main phase of the study. We adjust these items by subtracting the mean evaluation of the current product based on the 21 items from the follow-up study and then adjusting the difference by multiplication with β (quantile ranks follow $W_{\text{MSA}} = \sum_k \hat{\beta}(x_{\text{F}_k} - \bar{x}_{\text{C}})$). As can be seen in Figure 11, the improvement is again large, increasing the hit-rate by 21%.

Robustness Checks

We conducted two robustness checks to investigate the assumption of exchangeability of scale point across attributes and the minimum data requirements to calibrate the model. First, we test the assumption that a rating scale point is worth the same regardless of the attribute. This assumption allows us to apply β

Figure 12 (Color online) Split-Half Comparison of Posterior Distribution of $\ln \bar{\beta}_i$

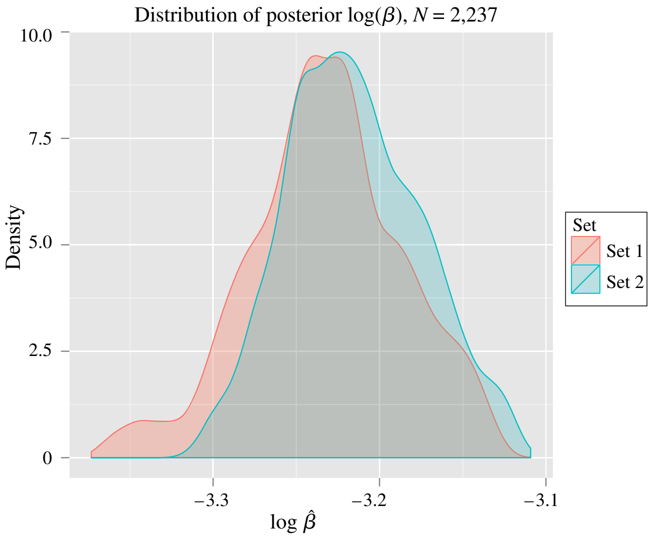


Table 5 Robustness Check: Validation Results with Reduced Number of Attributes

	WTP-level (%)					
Attributes	< \$2.50	\$2.50	\$2.75	\$3.00	\$3.25	\$3.50+
21	73	13	57	44	33	56
20	72	15	56	44	34	55
19	72	16	55	44	34	57
18	71	14	56	43	31	54
17	73	14	55	44	29	53
16	71	14	54	44	35	55
15	70	14	56	43	30	56
14	71	13	55	44	33	53
13	72	13	57	42	33	55
12	73	12	54	42	29	54
11	69	9	52	44	32	54
10	70	16	54	42	30	51
9	70	13	55	41	27	50
8	69	12	52	43	26	48
7	69	13	53	40	25	49

to the ratings of all attributes, even those beyond the scope of the calibration set. We conduct a split-half test, where the first set of attributes comprises all odd numbered attributes, and the second set consists of all even numbered attributes. We estimate the model using each set and compare the distribution of posterior means of $\{\log(\beta_i)\}$. Both distributions overlap 87%, supporting the assumption of exchangeability (see Figure 12).

In the second robustness check, we gradually reduce the number of attributes used for calibrating the model from 21 to 7. In each step, we randomly sample k attributes for each respondent and calibrate the model. We randomly pick attributes because the share of differences varies by attribute (see Figure 4), yet these shares would be unknown when designing a new study using our procedure. At each step, we adjust the entire set of 21 attributes using the $\{\beta_i\}$ obtained from calibrating based on k attributes. Then we compute hit-rates comparable to those in Figure 10. From Table 5, we can see that the model is remarkably stable even with lower numbers of attributes.

6. Discussion

In this section we explore two uses of the monetized scale, i.e., (i) to identify important product attributes that are drivers of product value; and (ii) to determine whether the monetized data better relates to other analyses conducted by the firm.

Driver Analysis

Figure 13 displays the relationship between the 21 items in the survey and the alternative focal/test products investigated in our analysis. The top portion of the figure summarizes the relationship by calculation of the average of the raw evaluations for the test product ($\bar{x}_F$). The middle portion displays the average difference relative to the respondent's current product ($\bar{x}_F - \bar{x}_C$), and the bottom portion weighs the relative difference with the monetary coefficient, $\sum_{n=1}^{N}(x_{Fn} - x_{Cn})\beta_n/N$. An efficient way of identifying important drivers is to array them rectangularly and color code the summary statistics. We use the color yellow to indicate no difference (or overall average for the raw data); positive values are colored green and negative values are red. Shades are determined by quantile ranks.

Comparing the raw, differenced, and monetized summaries indicates that monetizing "evens out" the results and leads to less extreme statistics. We find, for example, that attribute 14 tends to be measured high for all brands; a comparison of the bottom portion of the figure indicates that differencing relative to a competitive product alters inferences about the importance of this attribute. Similarly, attribute 16 is seen to take on a penalizing influence in the bottom two summaries. In general, we observe a great shift from vertical banding of responses in the top part of Figure 13 to horizontal banding at the bottom, i.e., from attribute importance to product importance. The monetized adjustment to the data therefore results in a greater distinction among the test products, or legs, of the survey in addition to better predictive performance.

The additional advantage of weighting the difference by β is that greater weight is given to respondents who are willing to pay for improvements. This weighting can be seen to cause changes in a number of attributes, such as attributes 6 and 7, which were identified as having zero relative value for products 3 and 4 when taking raw differences, as compared with having negative values in the bottom table in Figure 13. An additional benefit of using the weighted summaries at the bottom of the figure is being able to directly interpret the entries as the value of the test product

Figure 13 **Driver Analysis**

Raw data

Attr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Prod 1	3.2	3.1	3.4	3.6	3.7	3.5	3.6	3.2	3.1	2.9	3.4	3.9	3.8	4.1	3.8	3.4	3.7	3.7	3.6	3.4	3.9
Prod 2	3.6	3.7	3.7	3.8	4	3.7	4	3.5	3.3	3.2	3.9	4.3	4	4.5	3.9	3.6	3.9	3.7	3.7	3.4	3.8
Prod 3	3.7	3.4	3.8	3.8	3.9	3.8	3.9	3.2	3.1	3	3.8	4.2	3.7	4.5	3.7	3.5	3.8	3.7	3.6	3.4	3.8
Prod 4	3.5	3.5	3.4	3.6	3.8	3.6	3.7	3.3	3.1	2.9	3.8	4.2	3.8	4.5	3.8	3.5	3.8	3.7	3.7	3.4	3.8
Prod 5	3.7	3.9	3.9	3.9	4	3.7	3.9	3.4	3.3	3.1	4	4.3	4	4.5	3.9	3.6	3.9	3.7	3.8	3.4	3.8
Prod 6	3.8	3.6	3.9	4	4	3.8	3.9	3.5	3.5	3.3	4	4.3	4	4.5	3.7	3.6	3.9	3.8	3.7	3.4	3.8
Prod 7	4	3.8	4.1	4.2	4.2	3.9	4	3.8	3.7	3.6	4	4.1	3.9	4.5	4	3.7	4	3.8	3.7	3.5	3.9
Prod 8	3.6	3.5	3.7	4.1	4.1	3.9	4	3.6	3.5	3.4	3.7	4	3.9	4.4	3.8	3.5	3.9	3.7	3.8	3.5	3.9
Prod 9	3.8	4	3.9	4.1	4.2	3.9	4.1	3.7	3.6	3.6	4	4.2	4.2	4.5	4	3.7	3.9	3.8	3.8	3.5	4
Prod 10	3.8	4	3.7	3.8	4	3.8	3.9	3.5	3.3	3.2	4	4.2	3.9	4.6	3.9	3.7	3.9	3.7	3.8	3.5	3.9
Prod 11	3.8	3.8	3.9	4	4.1	3.9	4	3.6	3.4	3.3	4	4.3	4	4.6	4	3.8	4	3.8	3.9	3.5	3.9
Prod 12	4.1	4	4	4.2	4.2	4	4.1	3.6	3.5	3.4	4.1	4.2	4.1	4.5	3.9	3.8	4	3.8	3.9	3.5	3.9
Prod 13	3.9	3.9	4	4.3	4.2	3.8	4	3.8	3.7	3.5	4.1	4.3	4.2	4.6	4	3.8	4	3.8	3.9	3.6	3.9

Difference focal-current

Attr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Prod 1	−0.7	−0.6	−0.6	−0.5	−0.4	−0.2	−0.2	−0.3	−0.4	−0.3	−0.6	−0.3	−0.2	−0.4	−0.3	−0.9	−0.4	−0.2	−0.4	0	−0.1
Prod 2	−0.3	0	−0.3	−0.2	−0.1	0.1	0.2	−0.1	−0.2	−0.2	−0.1	0.1	0.1	0	−0.2	−0.6	−0.2	−0.1	−0.3	0	−0.1
Prod 3	−0.3	−0.5	−0.3	−0.3	−0.3	0	0	−0.3	−0.4	−0.3	−0.3	0	−0.3	0	−0.4	−0.8	−0.5	−0.1	−0.5	−0.1	−0.1
Prod 4	−0.3	−0.2	−0.5	−0.3	−0.2	0	0	−0.2	−0.3	−0.3	−0.2	0.1	−0.1	0	−0.2	−0.7	−0.3	−0.1	−0.3	0	−0.1
Prod 5	−0.1	0.4	0	0	0.1	0.2	0.2	0	0	0	0	0.1	0.2	0	−0.2	−0.6	−0.2	0	−0.2	0.1	0
Prod 6	0	−0.1	0	0.2	0.2	0.2	0.2	0.2	0.1	0.2	0	0.1	0.2	0	−0.3	−0.5	−0.3	0	−0.2	0	0
Prod 7	0.1	0	0.3	0.3	0.3	0.3	0.2	0.3	0.4	0.4	0	0	0.1	0	−0.2	−0.5	−0.2	0	−0.2	0.1	0.1
Prod 8	−0.3	−0.1	−0.2	0.2	0.2	0.3	0.3	0.3	0.1	0.3	−0.2	−0.3	0.1	−0.1	−0.3	−0.7	−0.2	0	−0.1	0.1	0
Prod 9	−0.1	0.5	−0.1	0.2	0.2	0.2	0.2	0.1	0.1	0.2	0	−0.1	0.3	0	−0.2	−0.6	−0.3	−0.1	−0.1	0	0
Prod 10	0	0.4	−0.2	−0.2	0	0.2	0.2	0	−0.1	−0.1	0.1	0.1	0.1	0.1	−0.2	−0.4	−0.2	−0.1	−0.2	0	0
Prod 11	−0.2	0.2	−0.1	0	0.1	0.2	0.2	0.1	0	0	0.1	0.1	0.2	0.1	−0.1	−0.4	−0.1	0	−0.1	0.1	0.1
Prod 12	0.2	0.3	0.1	0.2	0.3	0.3	0.3	0.1	0	0.2	0.2	0	0.3	0	−0.1	−0.4	−0.1	0	−0.1	0	0
Prod 13	0.2	0.5	0.3	0.4	0.4	0.4	0.3	0.4	0.4	0.4	0.2	0.2	0.5	0.1	0	−0.3	0	0.1	0.1	0.3	0.1

Monetized scale—% of reference price

Attr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Prod 1	−6	−7	−4	−5	−4	−2	−2	−3	−4	−4	−6	−3	−2	−3	−3	−7	−4	−1	−4	−1	−1
Prod 2	−2	0	−2	−1	0	1	1	−1	−1	−1	0	1	1	0	−1	−4	−1	−1	−1	0	0
Prod 3	−4	−6	−4	−5	−4	−1	−2	−5	−6	−5	−4	−2	−4	0	−4	−7	−4	−2	−4	−1	−1
Prod 4	−2	−1	−4	−3	−2	−1	−1	−1	−1	−1	−2	1	0	0	−2	−6	−3	−1	−2	1	0
Prod 5	−1	2	0	0	1	1	2	1	0	0	0	1	1	0	−1	−3	−1	0	0	1	0
Prod 6	0	−1	0	2	1	2	1	2	1	1	1	0	2	1	−2	−3	−2	0	−1	1	0
Prod 7	1	0	2	3	2	2	2	3	3	3	0	−1	1	−1	0	−2	0	1	−1	1	1
Prod 8	−2	0	−1	3	2	3	3	3	2	3	−1	−2	1	−1	−2	−4	−1	0	−1	1	1
Prod 9	1	6	2	4	4	3	4	2	2	3	1	0	4	0	−1	−3	−1	0	0	1	1
Prod 10	0	2	−1	−2	0	1	1	0	−1	−2	0	1	0	1	−1	−2	−2	0	−1	0	0
Prod 11	−1	2	0	0	1	1	1	1	0	0	0	1	1	1	0	−2	−1	0	0	1	1
Prod 12	1	1	1	1	2	2	1	0	0	1	1	0	2	0	−1	−3	−1	0	−1	0	0
Prod 13	2	6	4	5	4	3	3	4	4	5	3	2	5	1	1	−2	0	1	1	3	2

on a particular attribute in terms of a percentage of its price. Given the base price of $3, an entry of “2” translates to an expected value of $0.06.

Because responses are weighted by their monetary relevance, this raises the question of the effective sample size. Results from the analysis of adjusted data should not be dominated by just a handful of respondents. There are a number of respondents with small β_i (Figure 8) whose responses would be significantly down-weighted. The heterogeneity of weights implies

Table 6 Effective Sample Size

Group	1	2	3	4	5	6	7	8	9	10	11	12	13
Actual	152	193	164	156	171	167	173	171	166	174	185	182	183
Effective	73	91	54	73	75	67	71	70	55	67	79	69	74

Table 7 Share of Respondents with Higher WTP Given Higher Rating

Attribute	Percentage of purchasing at higher price
A1	38.3
A2	38.8
A8	31.5
A9	33.1
A10	33.0

that the effective base is smaller than the actual number of respondents. In effect, each of the i respondents is given a normalized weight $w_i = \beta_i / \sum_{i=1}^{N} \beta_i$. Using these weights, we can compute the effective base $b = 1/\sum_{i=1}^{N} w_i^2$ as defined in the survey sampling literature (Williams 2008). Effective bases are separately computed for each group. Overall, the effective base is 59% lower than the actual sample size (see Table 6). While this loss of effective sample size is substantial, sufficient sample size remains for inference.

Integrating Monetized Scales Into Existing Methods of Analysis

We expect that, if a monetized scale adjustment provides a purer interpretation of ratings data, then the adjusted data should be better related to other analyses conducted by the firm. In our case, the firm makes extensive use of Bayesian Belief Networks (BBN)[2] to estimate the effect of numerous attributes on measures such as WTP.

An important aspect of any scale use adjustment method is the capability to produce adjusted scores that can be used with existing and established methods of analysis. To evaluate our adjustment procedure, we used the WTP metric as the target variable in the BBN models. Because WTP is external to the rating scale, adjusting the rating scale predictor variables is not guaranteed to improve fit and predictions. An improvement in fit can thus be considered as validation of our method.

Use of adjusted data increases the overall accuracy of predicting WTP from 43.8% to 55.5%. In addition, the coefficients are different. Figure 14 provides a comparison of coefficients based on the raw and monetized ratings. The company considers standardized total effects, which consider direct and indirect effects of the explanatory variable on the target. The coefficients are standardized by dividing by the standard deviation of the target variable. It is common to select attributes with the largest total effects and consider them as "key drivers" that receive the most attention. We find that predictive accuracy is substantially enhanced using the monetized ratings. The adjusted ratings enable product developers to better focus on the most relevant attributes.

A key objective of our study, as identified by the firm, was to understand the relative importance of attributes 1, 2, and 8–10. The two driver analyses reported in Figure 14 provide contrasting results. Results from the monetized ratings suggest that attributes 1 and 2 are more important, while results from the analysis using raw ratings identify attributes 8–10 as more important. To investigate which attributes are most important, we compute the proportion of respondents with higher WTP given a higher evaluation for the test product versus their current product. Results indicate that higher ratings on attributes 1 and 2 are more strongly associated with increased WTP (see Table 7). Thus focusing on attributes 1 and 2 instead of 8–10 leads to higher percentages of respondents willing to pay more (at least by 5%), i.e., play a greater role in product choice.

Continuous and Discrete Attributes

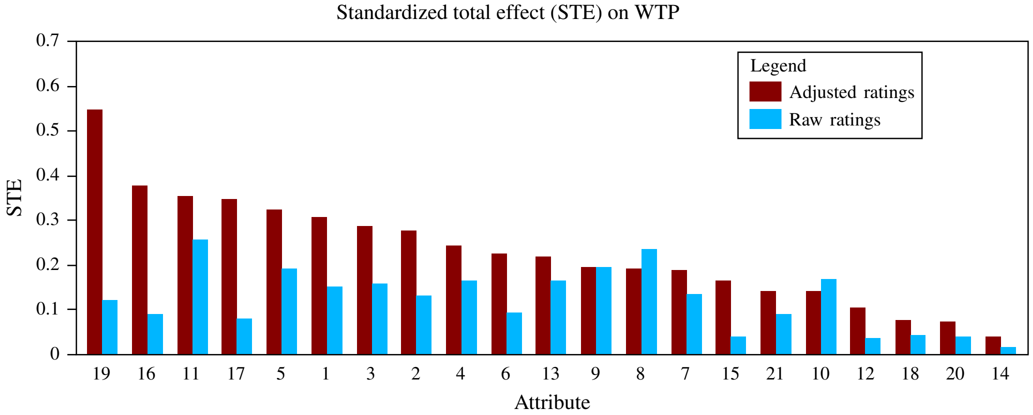
Our model requires all attributes to be assessed in terms of an ideal vector value scale "the more, the better." However, this limitation does not prevent us from studying discrete attributes. For instance, respondents can rate "color" on a value scale because actual colors are known to the company conducting the study. This reasoning also applies to attributes that are continuous in nature. For instance, when rating the "foam amount" of the test product in our study, engineers can link the prototype-specific value assessments to the specific foam amounts put into those prototypes. Figure 13 helps product engineers to understand consumer response to different formulations.

7. Concluding Remarks

This paper presents a new approach to scale adjustment based on the use of auxiliary data to monetize the value of a scale point difference. We develop the scale adjustment in the context of a product experience survey for a packaged goods manufacturer. The survey involves the evaluation of a test product with a large number (>50) of fixed-point rating scales; auxiliary data asks respondents to evaluate a subset of items of the original survey for their current product, and to choose between their current product and a hypothetical product at different prices that are the same except for one item that takes on the value of the test product.

[2] BayesiaLab (version 5.2), http://www.bayesia.us.

Figure 14 (Color online) BBN Results



The choice exercise is a binary "one-of" study in which one attribute of a product is changed at a time.

We find that the auxiliary data leads to a scale adjustment with better predictive properties than existing methods that difference and standardize the data. Our survey includes a number of holdout predictive tasks that allowed us to determine a respondent's WTP for the test product without use of a model. Our monetized scale adjustment leads to an improvement in predicting WTP by approximately 20% relative to the firm's current method of assessment.

Use of the monetized scale in driver analysis leads to more interpretable assessments of important attributes. We find that inferences about attribute importance change dramatically when measured relative to a respondent's current product, and that additional changes occur when the relative ratings are weighted by the monetary coefficient. Converting the ratings data to a monetized scale provides a direct interpretation of the monetary benefits of attribute improvement. Also, the monetized ratings are more predictive using other methods of analysis used by the firm.

An advantage of our proposed approach is that results do not depend on any redundancy in the attributes of the conjoint study. Traditional methods of scale adjustment, which do not involve auxiliary data, are heavily influenced by the items used to make the adjustment. That is, if the respondent does not have differentiated views about the scale items, because of lack of knowledge or lack of interest in providing accurate ratings, then the resulting scale adjustment will be minimal. By contrast, our use of auxiliary data conditions on evaluations of the focal test and competitive products, and queries respondents about the value of one attribute at a time. Equal evaluations of the two products provides no information about the scale adjustment parameter; random evaluations due to lack of interest are accommodated through the estimated scale parameter σ in the model. Thus, our approach is not affected by use of redundant items, and we show that the resulting scale adjustment can be successfully applied to all items in the survey, even when only a subset of items are used to estimate the scale adjustment (β) coefficient.

A strength of our method is that the use of focal test and competitive product evaluation alleviates the need to establish a scale origin. A disadvantage is that scale origins are sometimes useful for analysis, particularly for product categories that are not necessary goods. Our adjustment is useful for investigating changes to a product's formulation and determining relative preferences in a product category. It is not useful for understanding drivers of primary demand, purchase incidence in a product category or for understanding the drivers of preferences across product categories. We leave this issue for future research. In addition, our model focuses on demand-side information provided by respondents and does not include supply-side information about the costs of offering specific features. Although our model allows firms to assess the benefit, in monetary terms, of specific attributes, additional research is needed to fully integrate supply-side factors.

We encourage future studies that include additional information about the respondents' current products. If the set of current products is small, technical information may be obtained and compared. Moreover, it would be useful to anchor prices at the actual price of the household's current product. However, some consumers may not know the price of their current product or its exact name. Moreover, it would be interesting to track the respondents' actual purchases after the study to further validate our model.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/mksc.2016.0980.

Acknowledgments

The authors thank Michael Thompson for providing helpful feedback and comments.

Appendix

The likelihood specification is similar to a standard Logit specification with a fixed price coefficient but flexible error scale (Sonnier et al. 2007). The probability of observing j choices between the focal and current attribute level for respondent i is

$$\Pr(Y_i \mid \beta_i, \sigma_i, x_{\mathrm{F}_i}, x_{\mathrm{C}_i}, k_i) = \prod_j \{\pi_{ij}^{Y_{ij}} + (1-\pi_{ij})^{1-Y_{ij}}\},$$

where, following Equation (4)

$$\pi_{ij} = \frac{\exp[((x_{\mathrm{F}_{ij}} - x_{\mathrm{C}_{ij}})\beta_i - \ln k_{ij})/\sigma_i]}{1+\exp[((x_{\mathrm{F}_{ij}} - x_{\mathrm{C}_{ij}})\beta_i - \ln k_{ij})/\sigma_i]}.$$

We define

$$\Theta = (\{\beta_i\}\{\sigma_i\})^T.$$

Individual parameters are assumed to be distributed multivariate log-normal

$$\ln(\Theta_i) \sim N(\mu, \Sigma),$$

where

$$\mu \sim N(\bar{\mu}, A^{-1}),$$
$$\Sigma \sim IW(\nu, V_0).$$

We assumed weakly informative priors, $\bar{\mu} = (-3, -3)$ and $A^{-1} = 1/3I_2$ or $A^{-1} = 10I_2$, respectively. The choice of A^{-1} had little impact on the results. We get similar results assuming a conditional prior on μ, with $a^{-1} = 10$ or $a^{-1} = 100$

$$\mu \mid \Sigma \sim N(\bar{\mu}, \Sigma a^{-1}).$$

Including covariates in the upper level model is straightforward. For instance, the `rmultireg` function from the package `bayesm` in R can be used, which also assumes a conditional prior on μ.

The MCMC algorithm is straightforward:

Step 1. For each respondent i we apply the Metropolis–Hastings algorithm to draw from the posterior distribution of the adjustment parameter β_i and error scale σ_i, i.e., $(\beta_i, \sigma_i) \mid \mu, \Sigma$. The acceptance ratio is

$$\alpha_i = \frac{\Pr(Y_i \mid \beta_i^c, \sigma_i^c, \ldots)}{\Pr(Y_i \mid \beta_i, \sigma_i, \ldots)} \frac{\phi(\ln \Theta_i^c \mid \bar{\mu}, \Sigma)}{\phi(\ln \Theta_i \mid \bar{\mu}, \Sigma)},$$

where ^{c} denote proposal from the random walk algorithm.

Step 2. Gibbs sampling can be used for the upper level parameters

$$\mu \mid \Theta, \quad \Sigma \sim N(\tilde{m}, n\Sigma^{-1} + A),$$
$$\Sigma \mid \Theta \sim IW(\nu_0 + n, V_0 + S),$$

where

$$\tilde{m} = (n\Sigma^{-1} + A)^{-1}(\Sigma^{-1}\Theta\mathbf{1} + A\bar{\mu}),$$
$$S = (\Theta - \bar{\Theta}\mathbf{1}_n)'(\Theta - \bar{\Theta}\mathbf{1}_n),$$

and $\bar{\Theta}$ is the average of β and σ in a given draw.

Similarly, for the conditional prior

$$\mu \mid \Theta, \quad \Sigma \sim N(\tilde{m}, \Sigma(n+A)^{-1}),$$

where

$$\tilde{m} = (n+A)^{-1}(n\bar{\Theta} + A\bar{\mu}),$$

and

$$S = (\Theta - \tilde{m})'(\Theta - \tilde{m}) + (\tilde{m} - \bar{\mu})'A(\tilde{m} - \bar{\mu}).$$

References

Allenby GM, Rossi PE (1998) Marketing models of consumer heterogeneity. *J. Econom.* 89(1):57–78.

Allenby GM, Brazell JD, Howell JR, Rossi PE (2014) Economic valuation of product features. *Quant. Marketing Econom.* 12(4): 421–456.

Bacon L, Lenk P (2012) Augmenting discrete-choice data to identify common preference scales for inter-subject analyses. *Quant. Marketing Econom.* 10(4):453–474.

Baumgartner H, Steenkamp J-BEM (2001) Response styles in marketing research: A cross-national investigation. *J. Marketing Res.* 38(2):143–156.

Böckenholt U (2004) Comparative judgments as an alternative to ratings: Identifying the scale origin. *Psych. Methods* 9(4): 453–465.

Büschken J, Otter T, Allenby GM (2013) The dimensionality of customer satisfaction survey responses and implications for driver analysis. *Marketing Sci.* 32(4):533–553.

Chandukala SR, Kim J, Otter T, Rossi PE, Allenby GM (2008) Choice models in marketing: Economic assumptions, challenges and trends. Chandukala SR, Kim J, Otter T, Rossi PE, Allenby GM, eds. *Foundations and Trends in Marketing*, Vol. 2 (Now Publishers Inc., Hanover, MA), 97–184.

Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intelligence* PAMI-6(6):721–741.

Green PE, Rao VR (1971) Conjoint measurement for quantifying judgmental data. *J. Marketing Res.* 8(3):355–363.

Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57(1):97–109.

Johnson TR (2003) On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika* 68(4):563–583.

Krosnick JA (1991) Response strategies for coping with the cognitive demands of attitude measures in surveys. *Appl. Cognitive Psych.* 5(3):213–236.

Luo L, Kannan PK, Ratchford BT (2008) Incorporating subjective characteristics in product design and evaluations. *J. Marketing Res.* 45(2):182–194.

Meade AW, Craig SB (2012) Identifying careless responses in survey data. *Psych. Methods* 17(3):437–455.

Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E (1953) Equation of state calculations by fast computing machines. *J. Chemical Phys.* 21(6):1087–1092.

Rossi PE, Gilula Z, Allenby GM (2001) Overcoming scale usage heterogeneity: A Bayesian hierarchical approach. *J. Amer. Statist. Assoc.* 96(453):20–31.

Sonnier G, Ainslie A, Otter T (2007) Heterogeneity distributions of willingness-to-pay in choice models. *Quant. Marketing Econom.* 5(3):313–331.

van Rosmalen J, van Herk H, Groenen PJF (2010) Identifying response styles: A latent-class bilinear multinomial logit model. *J. Marketing Res.* 47(1):157–172.

Williams RL (2008) Effective sample size. Lavrakas PJ, ed. *Encyclopedia of Survey Research Methods* (Sage Publications, Thousand Oaks, CA).

Winkler JD, Kanouse DE, Ware JE (1982) Controlling for acquiescence response set in scale development. *J. Appl. Psych.* 67(5):555–561.

Ying Y, Feinberg F, Wedel M (2006) Leveraging missing ratings to improve online recommendation systems. *J. Marketing Res.* 43(3):355–365.