Accounting for Discrepancies between Online and Offline Product Evaluations

Daria Dzyabura
Stern School of Business, New York University, New York, NY 10012

Srikanth Jagabathula
Stern School of Business, New York University, New York, NY 10012

Eitan Muller
Stern School of Business, New York University, New York, NY 10012
Arison School of Business, The Interdisciplinary Center (IDC) Herzliya, 46101 Israel

January 2018

Accounting for Discrepancies between Online and Offline Product Evaluations

Abstract

Despite the growth of online retail, the majority of products are still sold offline, and the "touch and feel" aspect of physically examining a product before purchase remains important to many consumers. In this paper, we demonstrate that large discrepancies can exist between how consumers evaluate products when examining them "live" versus based on online descriptions, even for a relatively familiar product (messenger bags) and for utilitarian features. Therefore, use of online evaluations in market research may result in inaccurate predictions and potentially suboptimal decisions by the firm. Because eliciting preferences by conducting large-scale offline market research is costly, we propose fusing data from a large online study with data from a smaller set of participants who complete both an online and an offline study.

We demonstrate our approach using conjoint studies on two sets of participants. The group who completed both online and offline studies allows us to calibrate the relationship between online and offline partworths. To obtain reliable parameter estimates, we propose two statistical methods: a hierarchical Bayesian approach, and a *k*-nearest-neighbors approach. We demonstrate that the proposed approach achieves better out-of-sample predictive performance on individual choices (up to 25% improvement), as well as aggregate market shares (up to 33% improvement).

1. Introduction

Despite the rapid growth of online retail, the "touch-and-feel" experience of physically evaluating a product remains a significant driver of consumer purchase decisions. Physical evaluation drives purchase decisions in offline stores, which remain the predominant sales channel for many industries. In the first two quarters of 2017, online sales accounted for 8.7% of the $2.5 trillion in total US retail sales during this period (US Census Bureau 2017). A recent survey of 19K consumers by PricewaterhouseCoopers revealed that 73% of US consumers report having browsed products online, then purchased them in store. In the same survey, 61% of respondents cited being able to see and try out the item as the reason for buying in store (other reasons include delivery fees and having the item immediately).

With the prevalence of offline shopping, firms need to measure and predict consumer decision making in the offline channel. Yet conjoint analysis, widely used by firms to conduct market research, design products, and predict market shares, is nearly always conducted online, asking participants to rate or choose between product descriptions via computer. The implicit assumption is that the attribute partworths carry over from online behavior to offline behavior. However, consumers may weight attributes differently when evaluating a physical product than when reading a list of attributes on the computer. These two formats differ greatly in how they convey information to the consumer. Consumers might obtain more information about certain product attributes by evaluating the physical prototype; and they most commonly cite this reason for choosing to shop in physical stores. Additionally, an online description with the features presented in list form may render certain attributes more or less *salient* than when examining the product "live". Product features that are more salient tend to catch our attention and influence our decisions more than do less salient features. Behavioral factors are also at play that may affect how consumers process the information and integrate it into a decision in these two task formats.

In a conjoint experiment, we show that large systematic differences can exist between the weight that respondents give to product attributes online versus offline, even for utilitarian features such as a file divider and strap pad when evaluating messenger bags. We find that when the online task is performed first, the difference is larger than when the offline task is performed first, suggesting that much of the discrepancy is due to consumers' inability to obtain all necessary information about the product online, and learning more offline.

This discrepancy poses a problem, as conducting conjoint studies offline is significantly costlier than doing so online. We obtained price quotes from market research firms for a commercial offline conjoint study (to avoid peer effects, only one participant should be in the room examining products at a time). The costs involve payment to participants, the hourly rate of an experimenter (including salary, benefits, and overhead), and recruiting costs. The total comes to \$100-\$150 per participant, while online participants can be obtained for \$2-\$3 per participant. The firm therefore has access to very costly "accurate" data, and much cheaper data that are noisy but correlate with the accurate data. Given the large difference in costs, assuming budget constraint, our proposed solution is to split the budget between the two types of data.

The paper is organized as follows: The next section reviews relevant conjoint literature on using physical prototypes and on data fusion. Section 3 then describes Study 1, in which participants complete online and offline conjoint tasks in randomized order. Section 4 demonstrates superior predictive ability of offline choices when a separate group of online respondents' choices are fused with the data from Study 1. Section 5 reports an asymptotic variance analysis that calculates, for various cost-per-participant ratios, the precision with which partworths can be estimated for various sizes of online and offline populations. We conclude with implications and limitations of our work.

2. Related work

We contribute to the large body of literature on preference elicitation using conjoint analysis, first introduced to marketing by Green and Rao (1971). Since then, researchers have improved upon the basic methodology of asking respondents to rate, rank, or choose from among sets of products, with the goal of increasing the accuracy of the estimates of relative importance, or "partworths", of various product attributes. Netzer et al. (2008) provided a comprehensive review of recent developments in preference-measuring techniques, including conjoint analysis. They view preference measurement as comprising three components: (1) the problem the study seeks to address, (2) the design of the task and the data-collection approach, and (3) the specification and estimation of the model.

Under this framework, this paper specifically addresses the latter two steps (data collection, and estimation). Our proposed methodology for improving estimates of offline partworths consists of two components: (1) a data-collection method that measures both online and offline partworths for a set of respondents, and (2) a statistical data-fusion method that combines the online and offline data to estimate offline parameters. Each of these components builds on existing work, which we review here. We focus our review on two types of papers: those that propose to collect preference data using physical prototypes, and those that propose data-fusion techniques to combine conjoint data with other data.

Past research has demonstrated that using physical prototypes as part of the data collection process is feasible in a number of categories, and helps improve the validity of the preference elicitation. Srinivasan et al. (1997) advocated for the use of “customer-ready” prototypes rather than having consumers react to hypothetical product concepts, and demonstrated a discrepancy between evaluating descriptions and prototypes in the categories of citrus juicers, bicycle lighting systems, and travel mugs. Our task format is most similar to that of Luo et al. (2008), who asked respondents to rate prototypes on the likelihood of purchase, and used these ratings to infer attribute partworths, as part of a systematic approach to calibrating subjective product characteristics. More recently, a study by She and MacDonald (2013) exposed respondents to physical prototypes of toasters that either contained environmentally friendly features or not. They then measured attitude and choice in a consider-then-choose task. The manipulation of exposing respondents to the “trigger feature” did not induce respondents to either consider or purchase sustainable products more frequently. The key distinction between our work and the above body of literature is that we asked individuals to complete an online task in addition to evaluating physical prototypes offline.

The second step of our approach was to combine the online and offline data from the small set of respondents with online data of a larger set of respondents. Past research has developed methods for combining (or “fusing”) data from a conjoint study with another data source, such as aggregate market shares observed in the real market (Ben-Akiva et al. 1994; Feit et al. 2010; Orme and Johnson 2006; Swait and Andrews 2003). To combine preference data from two sources, most methods assume that the means of the partworths are similar in both datasets (e.g., Swait and Andrews 2003). Then parameters are estimated by combining the datasets using various statistical methods, such as incorporating market share data by introducing a constraint that requires parameter estimates to result in pre-specified market shares (Gilbride et al. 2008), or using the market shares as the prior in a Bayesian approach (Dzyabura and Hauser 2011). The approach proposed by Feit et al. (2010) provides more flexibility by linking preferences to consumer demographics, which are observed in both revealed (purchase) data and conjoint data. The demographic data allow for the population mean to differ between the two datasets.

All of the existing methods require the *same* individual to maintain the *same* individual parameter values across the two datasets. A key distinction of our work is that, depending on the order in which the respondent rated the products, the *same* individual may elicit *differing* resultant partworths in the online and offline datasets. We are able to merge the datasets by collecting both types of data from a set of consumers, which allows us to calibrate the mapping from online to offline preferences. Other work, such as that of Brownstone et al. (2000) and Bhat and Castelar (2003), has combined stated and revealed preference data from a panel of consumers, when *all* the respondents are observed in both datasets. Our approach allows offline data to be collected for only a subset of individuals, as collecting offline data is significantly costlier per respondent.

3. Online/Offline Discrepancy

Our first goal is to establish whether a significant discrepancy exists between the weight that consumers place on various product attributes when evaluating online versus offline[1]. To that end, we had a set of participants complete two conjoint tasks – one online and one offline – in randomized order. We found statistically significant differences between online and offline partworths both within and across subjects. The discrepancy is smaller when the offline task is done first than when the online task is done first.

[1] In section 3.3, we elaborate on the likely sources of this difference between the verbal or pictorial description of the online study vs. physical prototype offline presentation.

3.1 Study 1 design

For our studies, we used Timbuk2 messenger bags as the focus product. This product is a good example of our application because (1) these bags are often sold offline (as well as online), and (2) they are a familiar category, yet are infrequently purchased, such that we expected that many participants would not be familiar with some of the attributes and would therefore not have well-formed preferences. In addition, they are fully customizable through the firm's website, which allowed us to purchase bags for the offline study with the aim of creating an efficient experimental design.

Timbuk2's website offers a full customization option that includes a number of product features (http://www.timbuk2.com/customizer). We selected a subset of attributes that we expected to be relevant to the target population, and for which some uncertainty was likely to exist on the part of consumers and respondents. To make the study manageable, we reduced the number of levels of some of the features. We thus have the following six attributes for the study:

- Exterior design (4 options): Black, Blue, Reflective, Colorful
- Size (2 options): Small (10 x 19 x 14 in), Large (12 x 22 x 15 in)
- Price (4 levels): $120, $140, $160, $180
- Strap pad (2 options): Yes, No
- Water bottle pocket (2 options): Yes, No
- Interior compartments (3 options): Empty bucket with no dividers, Divider for files, Padded laptop compartment

We treat price as a continuous variable in the estimation, and have a total of 13 discrete attribute levels for the rest of the attributes. We recruited respondents through a university subject pool, and paid them $7 for completing both tasks, which together took 25 minutes on average. To ensure incentive compatibility and promote honest responses, the experimenter told participants that they would be entered in a raffle for a free messenger bag. Were they to win, their prize would be a bag configured to their preferences, which the researchers would infer from the responses they provided in the study. This chance of winning a bag provided an incentive to participants to take the task seriously and respond truthfully with respect to their preferences (Ding 2007; Hauser et al. 2010). We followed the instructions used by Ding et al. (2011) and told participants that, were they to win, they would receive a messenger bag plus cash, a combined value of $180. The cash component eliminates incentive for participants to provide higher ratings for more expensive items in order to win a more valuable prize.

Each participant was asked to complete an online conjoint task and an offline conjoint task. We used two conditions: subjects either completed the online task first followed by the offline task (Condition 1), or vice versa (Condition 2). We next describe the details of both tasks.

Conjoint task. We used a ratings-based task in which respondents rated each bag on a 5-point scale (Definitely not buy; Probably not buy; May or may not buy; Probably buy; Definitely buy). Using the D-optimal study-design criterion (Huber and Zwerina 1996; Kuhfeld, Tobias, and Garratt 1994), we selected a 20-product design that has a D-efficiency of 0.97. The reason a ratings-based task is preferable for offline conjoint is to keep the cost of the study reasonable. The cost of offline conjoint studies is affected not only by respondent time, but also by the number of physical prototypes that need to be created, which is not a factor in online conjoint. In our setting, conducting a choice-based conjoint (CBC) offline would require 75 distinct physical prototypes[2] (with each bag costing around $150), instead of the 20 we required for the ratings-based task design. Moreover, the task would require the researcher and the respondent to move between 20 displays of four prototypes each, potentially making the task tedious and the collected data prone to error. We use the ratings-based format in the online task as well, in the interest of keeping the tasks as similar as possible regarding all aspects other than online/offline.

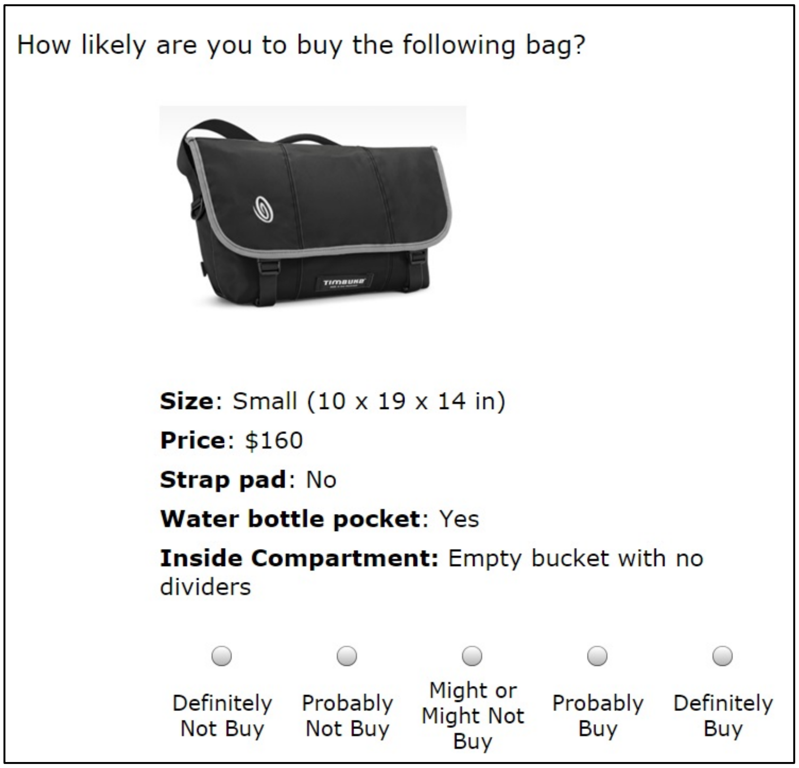
[2] Computed from Sawtooth: 20 choices among four profiles each, so as to obtain standard errors below 0.05. Note that this number could be somewhat reduced by appropriately constraining the choice design, while maintaining reasonable design efficiency, though lower than the one chosen here.

Because CBC is the more prevalent format, and because choices arguably have higher external validity, we demonstrate in Section 4 how results obtained from our data can be fused with an online CBC dataset to make predictions about participants' offline choices.

Online task. The online task was conducted using Sawtooth software. The first screens walked the participants through the feature descriptions one by one. Next, they were shown a practice rating question and were informed that it was for practice and that their response to it would be discarded. The screens that followed presented a single product configuration, along with the 5-point scale, and one additional question that was used for another study. Participants could go back to previous screens if they wished, but could not skip a question. Figure 1a shows a sample screenshot of the online task. The online portion took 10 minutes to complete on average.

Offline task. To ensure that participants could not see the bags during the online study, we conducted the offline task in a room separate from the computer lab in which the online task was conducted. This task was done individually, one respondent at a time in the room, to avoid a contagion effect. The bags were laid out on a conference table, each with a card next to it displaying a corresponding number (indexing the item), and the bags were arranged in order from 1 through 20. The prices were displayed on stickers on Timbuk2 price tags attached to each bag. The experimenter first walked the respondents through all the features, showing each one on a sample bag. Figure 1b shows the conference room display of the offline task. The offline portion took 15 minutes to complete on average.

Figure 1a: Sample online conjoint screen shot



How likely are you to buy the following bag?

Size: Small (10 x 19 x 14 in)

Price: $160

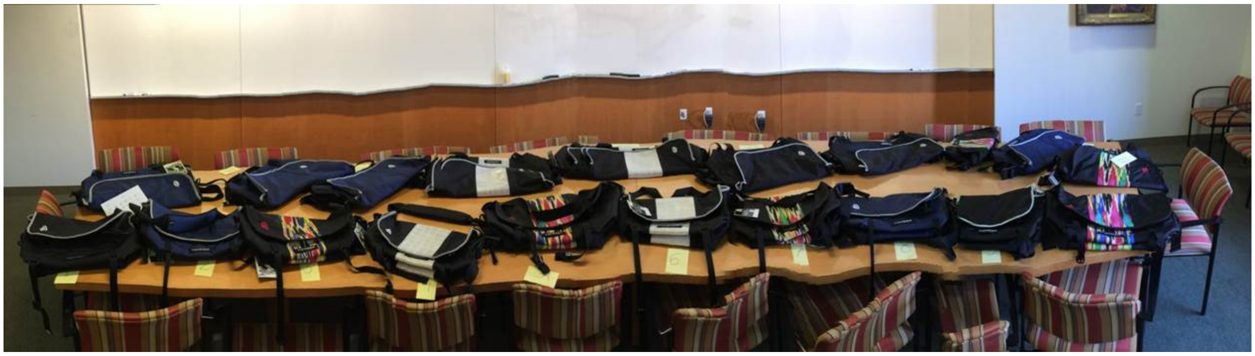
Strap pad: No

Water bottle pocket: Yes

Inside Compartment: Empty bucket with no dividers

○ Definitely Not Buy ○ Probably Not Buy ○ Might or Might Not Buy ○ Probably Buy ○ Definitely Buy

Figure 1b: Offline task room setup



We next describe the results of the study and demonstrate the discrepancy between partworths participants use when evaluating products online and offline.

3.2 Comparison of online and offline partworths

We begin with the online-first condition, which consisted of 122 participants. We assume a standard linear-in-attributes utility function. The categorical attributes are dummy coded, using one level of each category as a baseline; price is captured as a linear attribute. To capture consumer heterogeneity, we fit a linear mixed effects (LME) model to the ratings data, (abstracting away from any scale-usage heterogeneity):

$$u_{ijt} = \beta_{i0t} + \sum_{k=1}^{K} \beta_{ikt} x_{jk} + \epsilon_{ijt}, \tag{1}$$

$$\beta_{ikt} = \mu_{kt} + \delta_{ikt}.$$

In Equation (1), u_{ijt} is the rating by participant i of product j for task $t \in \{on, off\}$ with $t = on$ denoting the online task and $t = off$ denoting the offline task, β_{ikt} is the partworth that participant i assigns to feature *k* during task t, and β_{i0t} is the intercept. Product j is represented by its (*K*) attribute levels x_{jk}. The random component ϵ_{ijt} is assumed to follow a normal distribution, $\epsilon_{ijt} \sim N(0, \sigma_t^2)$; and the vector $\boldsymbol{\delta}_i = [\delta_{i1}, \dots \delta_{i,K}]$ follows a normal distribution with mean 0 and covariance matrix $\boldsymbol{\Sigma}_\mathrm{t}$ that is diagonal.

Table 1 reports the attribute fixed effects μ_k, estimated when the model in (1) is fit *separately* to the online and offline datasets. Standard errors are reported in parentheses.

Table 1: Mean population partworths ($\boldsymbol{\mu}$), online-first condition (standard errors in parentheses)

Attribute	Level	Online Partworth ($\boldsymbol{\mu_{k,on}}$)	Offline Partworth ($\boldsymbol{\mu_{k,off}}$)
Exterior design	Reflective	-0.31 (0.07)	-0.60 (0.09)
	Colorful	-1.06 (0.09)	-0.71 (0.10)
	Blue	-0.22 (0.06)	-0.11 (0.06)
	Black		
Size	Large	0.27 (0.05)	-0.31 (0.06)
	Small		
Price	\$120, \$140, \$160, \$180	-0.011 (8E-4)	-0.0075 (8E-4)
Strap pad	Yes	0.51 (0.05)	0.25 (0.05)
	No		
Water bottle pocket	Yes	0.45 (0.04)	0.17 (0.03)
	No		
Interior	Divider for files	0.41 (0.04)	0.52 (0.04)
	Crater laptop sleeve	0.62 (0.06)	0.88 (0.06)
	Empty bucket/No dividers		
Intercept		3.72 (0.12)	3.39 (0.13)

As we can see from Table 1, the magnitudes of the online and offline partworth differences are large for many attributes. The sign of the partworth of the ***size*** attribute, for example, even flips: Participants prefer the large size bag online, and the small size bag offline. The attributes strap pad and water bottle pocket carry much less weight offline than they do online.

3.3 Statistical tests to establish discrepancy

To formally compare the two sets of partworths, we use a nested-model likelihood-ratio test (LRT) to perform both within- and across-subject tests.

Within-subject test. We first test whether the online and offline partworths differ at the individual level. To do so, we estimate two models on the pooled online and offline data: one in which the online and offline parameters are constrained to be equal, and the other in which they are unconstrained. Specifically, the ***restricted*** model assumes that $\boldsymbol{\beta}_{i,on} = \boldsymbol{\beta}_{i,off}$ for all participants *i* and fits the following model to the pooled online and offline data, while the ***unrestricted*** model allows the participants to have differing partworths for each task:

(2) $$\text{Restricted: } u_{ijt} = \beta_{i0} + \sum_{k=1}^{K} \beta_{ik} x_{jk} + \epsilon_{ijt}, t \in \{on, off\}$$

The *unrestricted* model allows the participants to have differing partworths for each task

(3) $$\text{Unrestricted: } u_{ijt} = \beta_{i0t} + \textstyle\sum_{k=1}^{K} \beta_{ikt} x_{jk} + \epsilon_{ijt}, t \in \{on, off\}$$

but assumes that participant *i* samples the partworth vectors according to

(4) $$\begin{bmatrix} \boldsymbol{\beta}_{i,on} \\ \boldsymbol{\beta}_{i,off} \end{bmatrix} \sim \mathcal{N}\left(\begin{matrix} \boldsymbol{\mu}_{on} \\ \boldsymbol{\mu}_{off} \end{matrix}, \Sigma\right), \ \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{on} & \boldsymbol{\Sigma}_{on,off} \\ \boldsymbol{\Sigma}_{off,on} & \boldsymbol{\Sigma}_{off} \end{bmatrix},$$

where $\boldsymbol{\Sigma}_{on}$, $\boldsymbol{\Sigma}_{off}$, and $\boldsymbol{\Sigma}_{on,off}$ are diagonal. The estimates of the variance-covariance matrix **Σ** are reported in Appendix A. We constrain the covariance to only estimate covariance between the online and offline partworths between the same attribute level, e.g., blue online and blue offline. We use the LRT to test the null hypothesis[3] $\boldsymbol{\beta}_{i,on} = \boldsymbol{\beta}_{i,off}$ for all participants i. The log likelihoods of the restricted and unrestricted models are -6,969 and -6,691 respectively. There are 29 additional degrees of freedom in the unconstrained model, including 10 additional fixed-effects coefficients, and 19 additional covariance parameters. We are able to reject the null hypothesis because the LRT is significant ($p < 10^{-98}$).

Across-subjects test. One concern with the above within-subjects setup is that it may have led to a demand effect: If participants guessed that the researchers were looking for a difference between online and offline ratings, they may have felt compelled to change their decision rule in the offline task. To rule out this possibility, we used data from participants in Condition 2 (*N* = 40) who completed the offline task first. Comparing this group's offline ratings to the online ratings of the online-first group provides an across-subjects comparison of online and offline partworths, both of which came first for the respective group of participants. We test for significance again using the LRT. Because this study has an across-subjects design, we constrain only the fixed effects (μ_k) to be equal. In other words, we test the null hypothesis $\boldsymbol{\mu}_{on} = \boldsymbol{\mu}_{off}$. The constrained and unconstrained models' log likelihoods are -4,677 and -4,524 respectively, and the LRT again results in a significant difference, $p = 5 \cdot 10^{-48}$, allowing us to reject the null hypothesis. This finding suggests that when comparing the first task done by participants, unpolluted by any prior tasks, participants doing the online task use differing partworths than those doing the offline task.

[3] An alternative specification of the model would be $u_{ij,online} = \beta_{i0} + \Delta_{i0} + \sum_{k=1}^{K} \beta_{ik} x_{jk} + \sum_{k=1}^{K} \Delta_{ik} x_{jk} \cdot x_{online} + \epsilon_{ijt}$, where β_{ik} represents the offline attribute partworths, Δ_{ik} represents the bias due to the online format, and x_{online} is a binary variable that takes the value 1 in the online format and 0 in the offline format. In this specification, the null hypothesis can be stated more precisely as the population mean and variance parameters corresponding to Δ are zero.

3.4 Sources of online-versus-offline discrepancy

We have shown within and across subjects that a large, statistically significant discrepancy exists between partworths in online and offline task formats. While the main focus of our paper is on the discrepancy's consequences, we first discuss a possible theoretical framework that could explain the observed discrepancy. We do note, however, that our conjoint studies are not designed to isolate the underlying causes of the online-versus-offline discrepancy and as such, our theoretical framework provides only one possible explanation. Nevertheless, it demonstrates that the observed discrepancy is consistent with previous findings in the behavioral literature. We explore two mechanisms that have been studied in the consumer behavior literature that may be the source of this discrepancy: (1) *information* obtained from examining the products physically, and (2) inherent differences in attribute *salience* across the online versus offline formats.

Offline information gain. The first phenomenon that may be the cause of the discrepancy is the valuable information that consumers obtain about products by visually and physically examining them (Peck and Childers 2003). Learning through touch and feel occurs not only for inherently experiential attributes, such as color, size, and texture of the product, but also for utilitarian features. For instance, in the messenger bag study, examining physical products gives consumers information about just how padded the laptop compartment is, how much room it takes up in the bag, how easily accessible the water bottle pocket is, and so on.

Inherent attribute salience difference. Aside from additional information gained by physically examining products, the online and physical presentations also impart information to participants in differing formats, which may lead to behavioral biases. In the online channel, the attributes are presented in list form, whereas in the offline channel, the user sees the product as a whole. The attribute list representation may render certain attributes more or less salient to the user (Higgins, 1996). For example, attributes that are physically smaller, such as the water bottle pocket and the strap pad, are easy for the participant to miss when examining the bag physically, whereas the color and size of the bag are very noticeable. The phenomenon of consumers' choices being influenced by the format in which the information is presented to them is consistent with Bettman, Luce, and Payne's (1998) preference construction theory. Note that the attribute salience effect is inherent to each channel, and does not persist as the consumer moves from one channel to another. In this regard, the attribute salience effect differs from information gain, as the information obtained in the offline channel persists as the consumer moves to the online channel.

To assess how our theoretical framework explains the observed discrepancies, we further analyze the partworths in the two conditions to better understand the source of the discrepancy. These analyses are summarized in Figures 2a and 2b. Rather than testing behavioral theories, our goal is simply to demonstrate that the discrepancy between online and offline choice rules is consistent with that in previous work.

First, we note that in the online-first condition (Condition 1), both offline information gain and attribute salience influence the discrepancy between online and offline tasks for the same group of participants (Edge 1), while in offline-first condition (Condition 2) only attribute salience contributes to the discrepancy (Edge 3). This is because in Condition 2, any information gained from physically examining the products is obtained prior to completing the online task, and as mentioned above, the information obtained persists as the participant moves from the offline to the online task. The differences in the attribute salience between the channels exists in this condition as well because it is inherent to the format in which information was presented to the consumer.

Figure 2a. Mechanisms accounting for discrepancy between tasks

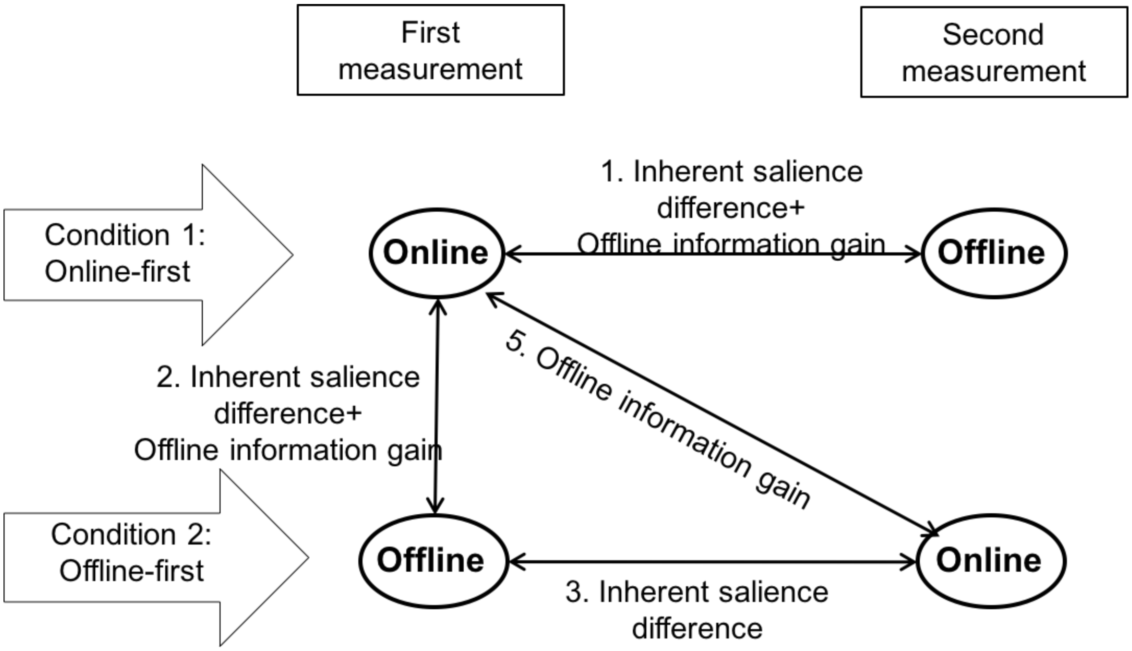
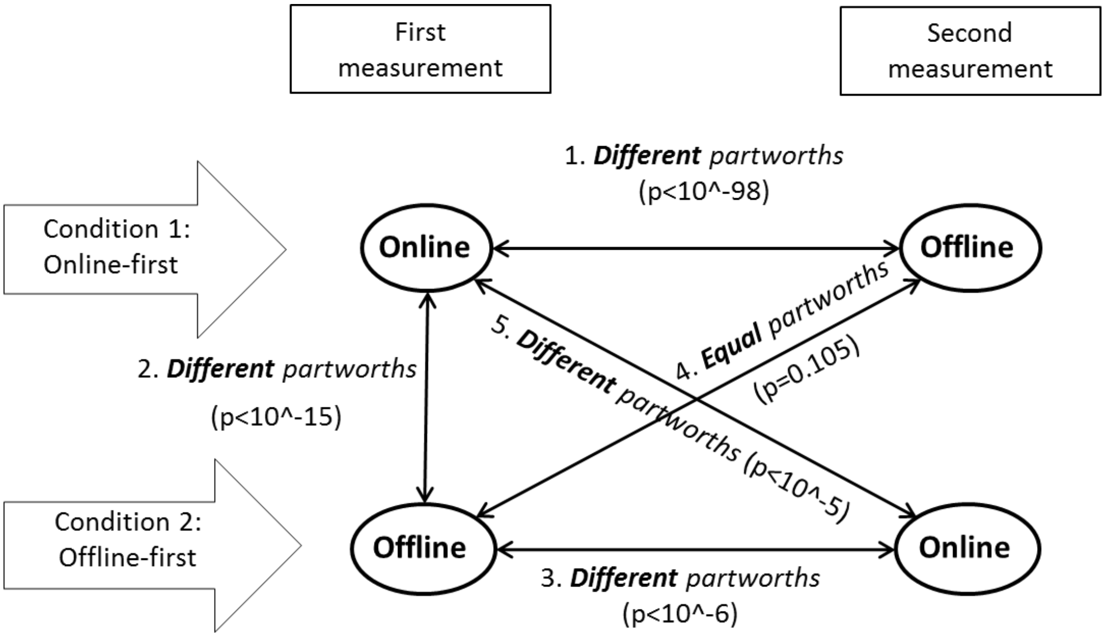

First measurement
Second measurement
Condition 1: Online-first
Condition 2: Offline-first
Online
Offline
Offline
Online
1. Inherent salience difference+ Offline information gain
2. Inherent salience difference+ Offline information gain
3. Inherent salience difference
5. Offline information gain

Figure 2b. Observed discrepancy between tasks


First measurement
Second measurement
Condition 1: Online-first
Condition 2: Offline-first
Online
Offline
Offline
Online
1. Different partworths (p<10^-98)
2. Different partworths (p<10^-15)
3. Different partworths (p<10^-6)
4. Equal partworths (p=0.105)
5. Different partworths (p<10^-5)

We find statistically significant differences in both cases (p < 10^-98 in Condition 1; and p = 10^-6 in Condition 2). In the offline-first condition, the magnitude of the difference between the partworths is much smaller than in the online-first condition. The online and offline partworths for the offline-first condition are presented in Table 2 (standard errors in parentheses).

Table 2: Mean population partworths ($\boldsymbol{\mu}$), offline-first condition

Attribute	**Level**	**Online Partworth** ($\boldsymbol{\mu_{k,on}}$)	**Offline Partworth** ($\boldsymbol{\mu_{k,off}}$)
Exterior design	Reflective	-0.25 (0.14)	-0.26 (0.16)
	Colorful	-0.26 (0.13)	-0.18 (0.14)
	Blue	-0.07 (0.11)	0.03 (0.10)
	Black		
Size	Large	-.15 (0.03)	-0.17 (0.07)
	Small		
Price	\$120, \$140, \$160, \$180	-0.012 (0.002)	-0.008 (0.002)
Strap pad	Yes	0.50 (0.09)	0.24 (0.08)
	No		
Water-bottle pocket	Yes	0.33 (0.07)	-0.005 (0.06)
	No		
Interior	Divider for files	0.65 (0.07)	0.62 (0.08)
	Crater laptop sleeve	1.01 (0.11)	1.15 (0.10)
	Empty bucket/no dividers		
Intercept		3.38 (0.21)	3.07 (0.25)

As compared with Table 1, we can see that for some attributes, very little discrepancy remains: For example, the size attribute is now weighted consistently in both conditions, consistent with our theory that the discrepancy we observed in the online-first condition was due to the participants being under-informed about the attribute based on the online description only. Other attributes, such as the water bottle pocket and strap pad, are weighted much higher online than they are offline, in *both* the online-first and offline-first conditions, suggesting the presence of channel-specific attribute salience effect. Our findings also suggest that exposing participants to physical prototypes prior to completing an online conjoint task (Feinberg, Kinnear, and Taylor 2012) produces partworth estimates that are closer to offline partworths, yet still leaving some discrepancy – possibly due to how the information is presented – such as the relative salience of the attributes online vs. offline.

We find additional evidence for information gain by comparing the two online tasks (Edge 5). The online task in Condition 2, which participants completed after the offline task, resulted in statistically significantly more differing (*p* = 10^-4) partworths than did the online task in Condition 1, which participants completed first. The restricted and unrestricted models' log likelihoods are –4,482 and –4,449 respectively. This difference is consistent with participants having obtained information by examining bags physically, and then applying this additional information to online product evaluation. Finally, we note that the online task did not appear to influence offline behavior, as we do not find statistically significant differences between the offline tasks between the two conditions (Edge 4). The log likelihood of the restricted model is -4,732, and that of the unrestricted model is -4,712. Note that since Edge 4 compares two offline tasks, there is no difference in attribute salience, nor is there additional offline information gain. The fact that we did not find statistically significant differences is consistent with our framework.

Discussion. We have shown that a discrepancy exists both within and across subjects. The discrepancy is reduced when the offline task is conducted first, suggesting that a large portion of the discrepancy is due to consumers being uninformed about some product attributes. Thus, if a firm conducts conjoint analysis online for a product that will be sold offline, the predictions it makes from the resulting partworths estimates are likely to be biased due to the online-versus-offline discrepancy. This estimate bias is a major issue if the aim is to make predictions about purchases made in the offline environment. Note that the information gain aspect of the discrepancy may be reduced by better informing consumers about the features, for example through more vivid descriptions or images.

The discrepancy between online and offline partworths has important consequences for a firm's decisions. For instance, Dzyabura and Jagabathula (2017) showed that if a firm sells products both online and offline, selecting the optimal product assortment requires knowledge of *both* online and offline partworths. Even when selling offline only, firms base both aggregate-level predictions, such as market shares, and individual predictions, such as segmentation or targeting decisions, on results of online preference elicitation. In these cases, ignoring the discrepancy can result in significant prediction errors. For instance, in the online-first condition, using the model estimated on participants' online ratings to predict their offline ratings results in an average RMSE of 1.56 (recall that ratings are on a 5-point scale). For comparison, the within-sample RMSE of using the offline ratings to predict the same ratings is 0.51. Therefore, it is important that firms correct the estimation bias to improve the accuracy of their decision-making.

We next address the issue of how the estimate bias can be corrected by supplementing an online conjoint study with a sample of respondents who complete both online and offline tasks.

4. Improving predictions of offline behavior

A straightforward solution to dealing with the systematic differences between consumers' online and offline preferences is to conduct offline conjoint studies rather than online studies featuring descriptions or images of products. In fact, past research advocates the use of physical prototypes to quantify the impact of subjective characteristics on consumers' purchase decisions (Luo et al. 2008; Srinivasan et al. 1997). But a large sample of respondents is necessary to obtain reliable parameter estimates, and conducting *large-scale* offline conjoint studies is logistically challenging and costly. Offline studies require the respondent to physically arrive at a location in order to participate, as opposed to the online studies, which can reach a large population of respondents on the web. This need for participants to be physically present increases the marginal cost of the study per respondent, rendering offline studies practically infeasible for all but a "small" number of respondents.

We propose a hybrid solution for improving the accuracy of offline preference estimates by supplementing a large online study with a small set of respondents who complete both an online and an offline task.

4.1 Correction techniques to predict individual-level offline partworths from online data

We now focus on the setting where our objective is to predict individual-level offline partworths. Concretely, we assume that we have asked a set I_{off} of respondents to complete an online conjoint task followed by an offline conjoint task; and another set I_{on} of respondents to complete only an online conjoint task. Instead of using the online partworth estimates for all the individuals in $I_{off} \cup \ I_{on}$, or using only the offline estimates for the individuals in I_{off}, we predict the offline partworths for the individuals in I_{on}, and use all offline partworth estimates for our decisions.

To predict the offline partworth estimates of the individuals in I_{on}, we propose two techniques: a Bayesian technique that we term Bayesian Inter-task Conditional Likelihood correction (Bayes ICL correction), and a *k*-nearest-neighbor (*k*-NN) technique. Both techniques rely on the observations of individuals in I_{off}, for whom we observe matched online and offline responses, to estimate the relationship between online and offline partworths. Based on this relationship, we predict the offline partworths using the online data from the individuals in I_{on}. The Bayesian ICL technique takes as input the raw observations of the individuals in I_{off} and

I_{on}, and produces as output the predicted offline partworths for individuals in I_{on}. The Bayesian technique relies on standard distributional assumptions. The k-NN technique is a machine-learning technique that makes no global distributional assumptions, but approximates the distribution locally. It takes as input the estimated partworth vectors $[\boldsymbol{\beta}_{i,on}, \boldsymbol{\beta}_{i,off}]$ for all individuals $i \in I_{off}$ and $\boldsymbol{\beta}_{i,on}$ $i \in I_{on}$; then outputs the predicted offline partworths for the individuals in I_{on}. The inputs for the k-NN technique can be obtained using any method. For the purposes of our empirical study, we use the estimates obtained from the Bayesian ICL method; see Section 4.2.

To test both methods on a holdout dataset, in addition to the results we obtained from Study 1, we conducted Study 2 with a group of 67 respondents who completed both an online and an offline task. We designed the study to mimic what a firm might do in practice. As in the first study, we asked each respondent to complete an online task, followed by an offline task. We use a choice task in this validation study, as that is the prevalent form of conjoint used in industry, and consumer choice data are more similar to what a firm would want to predict. The online portion was a traditional CBC task, consisting of 20 choice sets of four products each,[5] conducted using Sawtooth. We then presented the respondents with five choice sets of four products each in the offline environment.

The respondents' choices in the offline environment are the target variable we predict. We demonstrate that both the HB and k-NN corrections outperform the benchmark method of simply using the same respondents' online choice data to make predictions about their offline choices.

We now describe the two correction techniques we propose.

[5] We optimized the design of the study using the standard Sawtooth Complete Enumeration module to ensure efficient estimation.

Bayesian Inter-task Conditional Likelihood (ICL) correction

For this correction, we consider the following Hierarchical Bayesian (HB) model to describe the observations. We begin with the unrestricted linear model described in Equations (3) and (4) in Section 3. A key component of the model is its flexibility that allows the same consumer to assign differing partworths to the same feature when evaluating a product description online versus a physical product offline. To allow for this within-consumer discrepancy, the model associates the respondent's valuation of a given feature with two partworths, $\beta_{ikt}, t \in \{on, off\}$, where $\beta_{ik,on}$ is the utility partworths that respondent *i* applies to feature *k* online; and $\beta_{ik,off}$ is the utility partworth that s/he applies offline. That is, partworths vary by respondent, product feature, and task format (*on* for online and *off* for offline). The specification of the utility model in Equation (1) then becomes

$$u_{ijt} = \beta_{i0t} + \sum_{k=1}^{K} \beta_{ikt} x_{jk} + \epsilon_{ijt}, t \in \{on, off\}. \tag{5}$$

To accommodate a choice framework, we assume that the error terms $\epsilon_{ijt}\sim$ iid extreme value, instead of being normally distributed as in the linear model. We also let $c_t = 1, \ldots C_i^t$ index the choice tasks completed by respondent *i* in task format $t \in \{on, off\}$, X_{i,c_t} be the matrix containing sets of product attributes offered to respondent *i* in choice set c_t; and $X_{i,c_t,j}$ correspond to the *j*th row of matrix X_{i,c_t}, containing the attributes of the *j*th product in choice set c_t. The total number of products in respondent i's choice set c_t is J_{i,c_t}. Note that, because product attributes are categorial variables, online and offline attribute levels are coded as multiple levels of the same attribute. For example, if there are four colors, as in our data, in the online and offline data would be coded as one attribute with eight levels – four corresponding to online colors, and four corresponding to offline colors.

Finally, let y_{i,c_t} be respondent i's chosen product from set c_t. According to the utility specification in (5), the likelihood of observing choice y_{i,c_t} is given by:

$$[y_{i,c_t}|X_{i,c_t},\beta_{i,t}] = \frac{\exp\left(\beta_{i,t}\cdot X_{i,c_t,y_{i,c_t}}\right)}{\sum_{j=1,\ldots J_{i,c_t}}\exp\left(\beta_{i,t}\cdot X_{i,c_t,j}\right)}. \tag{6}$$

As above, we assume that the online and offline partworths are drawn from a joint multivariate normal distribution:

$$\begin{bmatrix}\boldsymbol{\beta}_{i,on}\\ \boldsymbol{\beta}_{i,off}\end{bmatrix} \sim \mathcal{N}\left(\begin{matrix}\boldsymbol{\mu}_{on}\\ \boldsymbol{\mu}_{off}\end{matrix},\Sigma\right), \qquad \boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_{on} & \boldsymbol{\Sigma}_{on,off}\\ \boldsymbol{\Sigma}_{off,on} & \boldsymbol{\Sigma}_{off}\end{bmatrix}. \tag{7}$$

Assuming that the respondents make choices according to the model specification in (6) and (7), we propose a Bayesian technique to estimate the parameters. Note that this method differs from the existing data-fusion techniques (e.g., Feit et al. 2010; Swait and Andrews 2003), as we allow the same individual i's $\beta_{ik,on}$ and $\beta_{ik,off}$ to differ from each other in both task formats, instead of constraining them to be equal. We estimate the population-level parameters as follows: Combining the individual- and population-level models gives us the likelihood of observing online and offline choices for all respondents:

$$\begin{aligned}
&[[\boldsymbol{y}|\mu_{on},\mu_{off},\boldsymbol{\Sigma},\boldsymbol{X}] = \\
&= \prod_{i\in I_{off}}\int_{\boldsymbol{\beta}_{i,on}}\int_{\boldsymbol{\beta}_{i,off}}\left(\prod_{c_{on}=1}^{C_i^{on}}[y_{i,c_{on}}|\boldsymbol{X}_{i,c_{on}},\boldsymbol{\beta}_{i,on}]\cdot\prod_{c_{off}=1}^{C_i^{off}}\left[y_{i,c_{off}}\middle|\boldsymbol{X}_{i,c_{off}},\boldsymbol{\beta}_{i,off}\right]\right. \\
&\left.\cdot\left[\boldsymbol{\beta}_{i,on},\boldsymbol{\beta}_{i,off}|\boldsymbol{\mu}_{on},\boldsymbol{\mu}_{off},\boldsymbol{\Sigma}\right]\right)d\boldsymbol{\beta}_{i,off}d\boldsymbol{\beta}_{i,on} \\
&\cdot\prod_{i\in I_{on}}\int_{\boldsymbol{\beta}_{i,on}}\prod_{c_{on}=1}^{C_i^{on}}[y_{i,c_{on}}|\boldsymbol{X}_{i,c_{on}},\boldsymbol{\beta}_{i,on}]\cdot[\boldsymbol{\beta}_{i,on}|\boldsymbol{\mu}_{on},\boldsymbol{\Sigma}_{on}]\,d\boldsymbol{\beta}_{i,on,}
\end{aligned} \tag{8}$$

where $\boldsymbol{y} = \{y_{i,c_t}\}$ is the set of all observed choices; and $\left[\boldsymbol{\beta}_{i,on}, \boldsymbol{\beta}_{i,off} | \boldsymbol{\mu}_{on}, \boldsymbol{\mu}_{off}, \boldsymbol{\Sigma}\right]$ is the probability of sampling the $\boldsymbol{\beta}$'s conditioned on the population-level parameters.

The model is estimated using a Bayesian approach. Because maximizing the likelihood expression in (8) over the population parameters is hard in general, we estimated the population parameters using the Bayesian framework. We used normal priors for $\boldsymbol{\mu}_{on}$ and $\boldsymbol{\mu}_{off}$ with mean 0 and variance 100, and followed Sawtooth software guidelines for setting the prior values for $\boldsymbol{\Sigma}_{off}, \boldsymbol{\Sigma}_{on}$. For $\boldsymbol{\Sigma}_{on,off}$, we set the prior value for the covariance of the same level of the same attribute in both the online and offline tasks to a positive value. The exact prior covariance is reported in Appendix C. As no closed-form expression for the posterior distributions of the parameters exists, we used a standard Gibbs sampler to generate samples of the unknown parameters ($\boldsymbol{\beta}_{on}, \boldsymbol{\beta}_{off}, \boldsymbol{\mu}_{on}, \boldsymbol{\mu}_{off}, \boldsymbol{\Sigma}$) iteratively (see Rossi et al. 2012). We then computed the individual- and population-level parameters by taking the average of 10,000 generated samples (after burning in the first 10,000 samples).

k-nearest neighbors (*k*-NN) correction

k-nearest neighbors (*k*-NN) is a popular data mining (meta-)algorithm, voted one of the top 10 data mining algorithms at the 2006 IEEE International Conference on Data Mining (ICDM) (Wu et al. 2008). It is a non-parametric method used for both classification and regression. It relies on the premise that 'similar' users behave similarly. In its most general form, the algorithm requires specification of a similarity function that produces a similarity score between pairs of users, a response variable of interest, and the number of nearest neighbors, *k*. Then, the algorithm predicts the response for a test user by outputting the weighted average of the responses of the *k* nearest neighbors in the training sample, as determined according to the given similarity metric. The weights may be chosen to be equal, proportional to the similarity scores between the test user and the corresponding neighbor, or optimized for prediction accuracy.

In our context, we begin with the premise that customers with similar values of online partworths will have similar values of offline partworths. Based on this premise, we approach the following prediction task: We are given both the offline partworth vector, $\boldsymbol{\beta}_{i,off}$, and the online partworth vector, $\boldsymbol{\beta}_{i,on}$, for each individual $i \in I_{off}$; and the online partworth vector $\boldsymbol{\beta}_{i,on}$ only for individuals $i \in I_{on}$. Our objective is to predict the offline partworth vector for all of the individuals in the set I_{on}. The given partworth vectors themselves could be estimated using any method. For the purposes of the empirical study presented below, we use the partworth estimates obtained from the HB method described above.

To predict the offline partworth of a respondent in I_{on}, we select the *k* nearest respondents in the set I_{off}, where the distance between two respondents is measured as the Euclidean distance between their respective online partworths:

$$(9) \qquad d(i,i') = \sqrt{\sum_{k=1,\ldots K} \left(\beta_{i,k,on} - \beta_{i\prime,k,on}\right)^2}.$$

For each respondent $i \in I_{on}$, we let S_i denote the set of *k* nearest neighbors from the set I_{off}. Then, we predict the respondent's offline partworth as:

$$(10) \qquad \widehat{\boldsymbol{\beta}}_{i,off} = \sum_{i' \in S_i} w_{i,i'} \boldsymbol{\beta}_{i',off},$$

where the weights $w_{i,i'} = 1/d(i,i')$ are chosen to be inverses of the distances between the corresponding individuals. Note that by construction we have $S_i \subset I_{off}$ and we are given $\boldsymbol{\beta}_{i',off}$ for all individuals $i' \in I_{off}$. Therefore, we can compute the expression in (10). The value of *k* is typically tuned using cross-validation, which is standard practice for model selection (Abu-

Mostafa, Magdon-Ismail, and Lin 2012). For the purposes of the empirical study, we picked *k* = 30 through 10-fold cross-validation. Specifically, we split the individuals in the set I_{off} into 10 (roughly) equal parts, and then train our model on data from 9 parts and make predictions for the individuals in the 10th part. Letting $I_{off,train}$ denote the individuals in the first 9 parts and $I_{off,validation}$ denote the individuals in the 10th part, we compute the prediction error

$$(11) \qquad Error = \sum_{i \in I_{off,validation}} \sum_{k=1,\ldots K} \left(\beta_{i,k,off} - \hat{\beta}_{i,k,off}\right)^2$$

where we compute $\widehat{\boldsymbol{\beta}}_{i,off}$ using Equation (10), but with neighborhood of individual *i* chosen as the *k* closest individuals from the set $I_{off,train}$, where the distance between individuals is measured as in Equation (9). We repeat the above process 10 times (folds) with each of the 10 parts used exactly once as the validation sample $I_{off,validation}$. We average the error across the 10 folds, and use the resulting average error as the proxy for the out-of-sample performance. We compute the average error for each value of the number *k* of neighbors from the set {10, 15, 20, 25, 30, 35, 40, 35}, and pick the value that resulted in the least average error. We found that *k* = 30 resulted in the least average error.

The *k*-NN approach is custom-built for making individual-level predictions. It is particularly well-suited to settings in which the population distribution is multi-modal (Wu et al. 2008). In this case, a Bayesian approach with a normal distribution assumption shrinks the all the respondents' individual-level partworths to a *single* population mean. Instead, the *k*-NN approach shrinks the partworths of different respondents to means of different subsets of respondents, and thereby better captures multiple modes within the dataset. Once the partworth vectors for the individuals in I_{on} and I_{off} are given, the *k*-NN does not make any parametric assumptions in predicting the offline partworths for the individuals in I_{on}. In this sense, it is a nonparametric method that is not designed to incorporate any prior information. Therefore, all else being equal, we expect it to perform better when the modeler has access to a larger training sample of respondents with offline partworths. We implemented the *k*-NN method using the sklearn.neighbors.KNeighborsRegressor function from Python scikit-learn (Pedregosa et al. 2011).

4.2 Performance of the Bayesian ICL and *k*-NN corrections

We now describe the empirical performance both in terms of prediction and decision accuracies of the proposed Bayesian ICL and *k*-NN corrections. To test the performance, we used data from two studies: first, the data on the respondents in Study 1, Condition 1, who did an online followed by an offline conjoint (N = 122), described above in Section 3; and second, the data from a set of respondents who completed an online CBC followed by an offline CBC as part of Study 2 (N = 67), described next.

Study 2

As in Study 1, we asked each respondent to complete an online task, followed by an offline task. Prior to completing the choice tasks, respondents first viewed a screen with instructions and then viewed each attribute description one by one, followed by a sample choice task that we did not use for estimation. For the offline task, we created five choice sets of four bags each, using the same 20 physical bags as in the first study. For each respondent, bags in the same choice set were placed next to each other, and each choice set was covered with fabric to avoid comparison to previous or subsequent choices, and to ensure that the respondents focused on the products in the present choice set. We also positioned the sets of bags such that two consecutive choice tasks were not next to one another; for example, Choice Set 1 was not next to Choice Set 2. We used this approach to help the respondent focus on the single choice set she was presented with and to refrain from comparing the bags to those in the choice sets she had just seen. The experimenter pointed out each of the features on a sample bag, and then uncovered one choice set at a time, in order from 1 to 5. Respondents circled their choices on a paper form. Completing this portion took respondents about 15 minutes.

To avoid idiosyncratic noise in the performance measures, we randomly assigned respondents to one of three groupings of the bags into choice sets. Although all respondents' choice sets were comprised of the same 20 bags, the bags were divided differently into five choice sets, resulting in 15 distinct choice sets total.

Prediction accuracy: Predicting offline choices using online data

We first assess the overall improvements in the accuracy of predicting offline choices of individuals in I_{on}, who completed Study 2. The data used for our computational study is summarized in Table 3. Our objective is to predict the choices in cell D in Table 3.

Table 3: Data used for predictive performance

	Study 1($N = 122$)	Study 2 ($N = 67$)
Online task	20 ratings A	20 choices B
Offline task	20 ratings C	5 choices D

We let I_{off} denote the set of individuals from Study 1 who completed an online task followed by an offline task (cells A and C in Table 3); and I_{on} denote the set of individuals who were part of Study 2 (corresponding to cells B and D in Table 3). Because respondents in I_{off} completed a ratings-based task, and the HB method assumes choice data, we first convert these ratings to a choice format. We adapted a procedure similar to the *rank-ordered logit model* (Beggs et al. 1981; Hausman and Ruud 1987), also known as the *exploded logit model* (Allison and Christakis 1994; Chapman and Staelin 1982; Punj and Staelin 1978) for ranking data.

Specifically, let u_{ij} denote the rating provided by participant i for bag j. We converted the ratings data into a ranked list by breaking ties based on the population averages. Let $\bar{u}_j$ denote $\frac{\sum_{i \in I_{off}} u_{ij}}{|I_{off}|}$, or the rating that product j received averaged over the entire population. Then, with each individual i, we associated ranked list δ_i that encodes the ranking in terms of pairwise comparisons, with δ_{ijl} taking the value 1 if the participants preferred j to l, and 0 otherwise. More precisely, $\delta_{ijl} = 1$ if $j = l$ or $u_{ij} > u_{il}$ or if $u_{ij} = u_{il}$ and $\bar{u}_j > \bar{u}_l$, and 0 otherwise.

We then converted the ranked list into "exploded" choice sets as follows: Fix an individual i. Let $(j_1, j_2, \dots, j_{20})$ be the ordered ranked list corresponding to δ_i, with the products ordered in decreasing order of preference. We "explode" this ranked list into 20 choice sets: $C_1 = \{j_1, j_2, \dots, j_{20}\}, C_2 = \{j_2, \dots, j_{20}\}, \dots, C_{20} = \{j_{20}\}$, where we successively remove the most preferred product from each choice set. The respondent's choice is the most preferred product from each set, so the choice y_{i,C_l} is equal to j_l for $l = 1, 2, \dots, 20$.

We compared three different methods for predicting the offline choices of participants in Study 2: the benchmark method, the Bayesian ICL correction, and the *k*-NN correction. The online benchmark method ignores all the data from the offline studies. Using the standard Bayesian techniques, the method estimates the expected online partworths $\boldsymbol{\beta}_{i,on}$ for each individual *i*, and sets $\boldsymbol{\beta}_{i,off,bench} = \boldsymbol{\beta}_{i,on}$. The Bayesian and *k*-NN techniques estimate $\widehat{\boldsymbol{\beta}}_{i,off,Bayes}$ and $\widehat{\boldsymbol{\beta}}_{i,off,kNN}$ as described above, using online data from participants in Study 2.

To assess the quality of the estimates, we compute two metrics on the held-out offline choices: individual log-likelihood and choice-set RMSE, defined as follows:

Individual log likelihood:

(12) $$\frac{1}{|I_{on}|}\sum\nolimits_{i\,\in\,I_{on}}\frac{1}{\left|C_i^{off}\right|}\sum_{c=1,\ldots C_i^{off}}\log\frac{\exp\left(\widehat{\boldsymbol{\beta}}_{i,off,method}\cdot X_{i,c,y_{i,c}}\right)}{\sum_{j=1,\ldots J_{i,c}}\exp\left(\widehat{\boldsymbol{\beta}}_{i,off}\cdot X_{i,c,j}\right)},$$

where c indexes the offline choice task of respondent i, and $y_{i,c}$ denotes the product chosen by the respondent in choice task c, $j = 1, \ldots J_{i,c}$ denote the products offered in choice task c, and $method \in \{bench, Bayes, kNN\}$. Of course, the higher the value of the likelihood, the better the method.

To compute the choice-set share RMSE, let $m_{j,c}$ denote the observed market share for product j and choice set c, and let $\widehat{m}_{j,c,method}$ denote the predicted market share, given by

(13) $$m_{j,c} = \frac{\left|i \in I_c : y_{i,c} = j\right|}{|I_c|},$$
$$\widehat{m}_{j,c,method} = \frac{1}{|I_c|}\sum_{i\in I_c}\frac{\exp\left(\widehat{\boldsymbol{\beta}}_{i,off,method}\cdot X_{i,c,y_{i,c}}\right)}{\sum_{j'=1,\ldots J_{i,c}}\exp\left(\widehat{\boldsymbol{\beta}}_{i,off,method}\cdot X_{i,c,j'}\right)},$$

where I_c is the set of respondents presented with choice set c. Then the choice-set share RMSE metric is given by

(14) $$\frac{1}{|C|}\sum_{c\in C}\sqrt{\sum_{j=1,\ldots J_c}\left(m_{j,c} - \widehat{m}_{j,c,method}\right)^2},$$

where C denotes the collection of all the choice tasks.

We report the performance of the methods (Bayesian ICL correction and *k*-NN correction), and two benchmarks, on both performance metrics in Table 4.

Table 4: Predictive performance

	Individual log likelihood	**Choice-set share RMSE**	**% change Individual log likelihood**	***p*-val**	**% change Choice-set share RMSE**	***p*-val**
Bayesian	-1.347	0.177	17.9%	0.001	**33.6%**	0.013
***k*-NN**	-1.191	0.267	**27.3%**	0.000	-0.7%	0.95
Benchmark	NA	0.246	NA		7.5%	0.62

(offline only)				
Benchmark (online only)	-1.640	0.266	NA	NA

In addition to the online benchmark mentioned above, we also add an offline benchmark that ignores the data from all the online studies. Since the offline benchmark ignores the data from cell B in Table 3, it can predict aggregated market shares only, and not individual-level partworths for individuals in I_{on}. Therefore, we report the choice-set share RMSE metric only for the offline benchmark. The prediction task for all four models is the offline choice data in study 2 (cell *D* in Table 3). The online-only benchmark uses the online data from both sets of participants, i.e., cells *A* and *B*. The offline-only benchmark uses the data in cell *C*. Therefore, we only compute aggregate level choice set shares for the offline-only benchmark, as we cannot make individual-level predictions. The two corrections (Bayesian and *k*-NN) use the online data from both sets of participants, and the offline data from the first group of participants, cells *A*, *B,* and *C* of Table 3. Significance is computed relative to the online-only benchmark using a two-tailed paired sample t-test, over the 67 individuals for the individual log likelihood metric, and 15 choice-sets for the choice-set share RMSE metric.

From Table 4, we observe that both proposed methods that use the offline data from the first study to estimate participants' offline partworths outperform the online-only benchmark, which simply uses their online partworths. We also compute aggregate-level predictions from the offline data. Note that the Bayesian method leads to a larger improvement in the aggregate measure, whereas the *k*-NN method leads to a larger improvement in the individual prediction, while demonstrating no significant improvement in the aggregate prediction. One reason for the difference in aggregate-level performance may be that the Bayesian estimation specifically estimates population-level parameters, as part of the model estimation. The *k*-NN method, on the other hand, does not have the population mean as an explicit model parameter that is estimated. The difference in individual-level performance is likely due to *k*-NN being particularly well suited to individual-level predictions when the population distribution is multi-modal (Wu et al. 2008), as discussed in section 4.1. Because the two methods are respectively better tailored to different predictive tasks, researchers can choose which method is appropriate depending upon whether individual-level or aggregate-level predictions are more important in their particular application.

5. Robustness and value of the offline information

In this section, we evaluate the robustness of our findings as well as the value of offline information and in particular, the role of online data in inferring offline parameters, and the relative information gain between offline and online data. We begin with a robustness evaluate our two correction methods, namely the Bayesian ICL correction, and *k*-NN correction , in terms of their prescriptive implications for optimal product lines.

5.1 Robustness test: Comparison of the revenue-maximizing subsets

We now investigate how products' revenue-maximizing subsets differ under differing methods of estimating utility partworths. The problem of finding revenue- or sales-maximizing subsets of products, also called the *Assortment Optimization* (AO) problem, has received much attention both in the marketing (Kohli and Sukumar, 1990) and the operations management (Kök et al., 2015) literature. It is aimed at helping a firm make the important decision of which products to carry in its stores. As it involves solving a computationally challenging set optimization problem, most existing work has focused on developing tractable techniques for addressing the computational challenge.

The critical inputs to these techniques are utility partworths, which are often estimated from a conjoint study, as we did here, or using secondary transaction-level data. These partworth estimates' accuracies crucially determine the accuracy of the assortment decision. We now showcase how the particular corrections to partworth estimates that we propose impact a firm's assortment decision. For that, we compute the revenue-maximizing product subsets of size four using the partworth estimates obtained from three different methods: the online-only benchmark, Bayesian ICL correction, and *k*-NN correction. To compute the revenue-maximizing subset for each method, let $\beta_{i,method}$ denote the estimated utility partworths for participant i in Study 2. For a subset *S* of bags, we compute the expected revenue under a particular method as follows:

$$(15) \qquad R^{method}(S) = \frac{1}{|I_{on}|} \sum_{i \in I_{on}} \frac{\sum_{j \in S} r_j \ \exp\left(\boldsymbol{\beta}_{i,method} \cdot \boldsymbol{X}_j\right)}{\sum_{j \in S} \exp\left(\boldsymbol{\beta}_{i,method} \cdot \boldsymbol{X}_j\right)}$$

where r_j denotes the price of product *j*, and I_{on} denotes the set of participants who completed Study 2. We then search over all possible subsets of size four from the 20 bags used in our two studies, to obtain the revenue-maximizing subset.

Table 5 presents the results. We observe from the table that the assortment decisions under the benchmark method and both corrections overlap in only two products, namely A and C. This marginal overlap suggests that the differences in the partworths can significantly alter the assortment decision, and thereby the firm's revenue and profit potential. In addition, we notice that the subsets obtained under both the corrections overlap in three products, namely A, B, and C, out of a possible four. This suggests that the corrections are robust, at least as far as the assortment decision is concerned.

Table 5: Revenue-maximizing subsets of size four under different methods*

Method	**Color**	**Size**	**Strap**	**Water bottle**	**Interior**	**Price**	**Product ID**
Benchmark (online only)	colorful	small	yes	yes	laptop	$180	A
	black	large	yes	yes	laptop	$140	C
	reflective	large	no	yes	divider	$180	D
	blue	large	yes	no	empty	$180	E
Bayesian correction	colorful	small	yes	yes	laptop	$180	A
	blue	small	no	yes	divider	$160	F
	black	small	no	no	laptop	$180	B
	black	large	yes	yes	laptop	$140	C
***k*-NN correction**	blue	small	yes	yes	divider	$140	G
	colorful	small	yes	yes	laptop	$180	A
	black	small	no	no	laptop	$180	B
	black	large	yes	yes	laptop	$140	C

* The subsets obtained under both corrections overlap in three out of four products (A, B, and C), indicating that the assortment decision is reasonably robust to the particular correction used.

5.2 Role of online data in inferring offline parameters

To obtain intuition on how online data helps improve offline estimates, we analytically illustrate the correction under a linear mixed-effects model. Under this model, it can be shown that, conditioned on the online parameters, the offline parameters are distributed as a multivariate normal random variable:

$$\boldsymbol{\beta}_{i,off} | \boldsymbol{\beta}_{i,on} \sim \mathcal{N}\left(\boldsymbol{\mu}_{i,off|on}, \boldsymbol{\Sigma}_{i,off|on}\right),$$

$$(16) \qquad \boldsymbol{\mu}_{i,off|on} = \boldsymbol{\mu}_{off} + \boldsymbol{\Sigma}_{on,off} {\boldsymbol{\Sigma}_{on}}^{-1} \cdot \left(\boldsymbol{\beta}_{i,on} - \boldsymbol{\mu}_{on}\right),$$

$$\boldsymbol{\Sigma}_{i,off|on} = \boldsymbol{\Sigma}_{off} - \boldsymbol{\Sigma}_{off,on} {\boldsymbol{\Sigma}_{off}}^{-1} \boldsymbol{\Sigma}_{on,off}.$$

We can rewrite the conditional distribution of $\boldsymbol{\beta}_{i,off}|\boldsymbol{\beta}_{i,on}$ as the sum of a deterministic component and a mean zero random component: $\boldsymbol{\beta}_{i,off}|\boldsymbol{\beta}_{i,on} = \boldsymbol{\mu}_{i,off|on} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \mathcal{N}\left(0, \boldsymbol{\Sigma}_{i,off|on}\right)$. Substituting the expression for $\boldsymbol{\mu}_{i,off|on}$ from Equation 16, we obtain the following relationship:

$$(17) \qquad \boldsymbol{\beta}_{i,off}|\boldsymbol{\beta}_{i,on} - \boldsymbol{\mu}_{off} = \boldsymbol{\Sigma}_{on,off}{\boldsymbol{\Sigma}_{on}}^{-1} \cdot \left(\boldsymbol{\beta}_{i,on} - \boldsymbol{\mu}_{on}\right) + \boldsymbol{\varepsilon}.$$

In Appendix B, we show that under some constraints on the structure of the covariance matrix $\boldsymbol{\Sigma}$, one can rewrite Equation (17) for individual attribute *k* as follows:

$$(18) \qquad \beta_{i,k,off|on} - \mu_{k,off} = \rho_k \frac{\sigma_{k,off}}{\sigma_{k,on}} \left(\beta_{i,k,on} - \mu_{k,on}\right) + \varepsilon_k,$$

where ρ_k is the correlation coefficient between the online and offline partworths $\beta_{i,k,off}$ and $\beta_{i,k,on}$. Here, we can see the extent to which individual-level offline and online partworths for an attribute are directly related: If an individual had a higher than average partworth for a particular feature *k* online, $\beta_{i,k,on} > \mu_{k,on}$, she will also have a higher than average partworth for the feature offline, assuming the online and offline partworths are positively correlated ($\rho_k > 0$ in Equation 18).

We can see from Equations (17) and (18) that if the correlation is zero (or close to zero), then $\beta_{i,k,on}$ is not a good predictor of $\beta_{i,k,off}$, and our conditional estimate of the individual offline partworth is simply the population mean. On the other hand, the *higher* the magnitude of the correlation between the online and offline partworths of the same attribute level $|\rho_k|$, the more precisely we can estimate the *individual*-level $\beta_{i,k,off}$ from $\beta_{i,k,on}$. In particular, if $\beta_{i,k,off}$ and $\beta_{i,k,on}$ are perfectly correlated, we obtain an estimate of $\beta_{i,k,on}$ simply by subtracting the difference between the two population means. For example, note that (from Table 1 and Appendix A) the feature "colorful" has online-offline *bias*, or high $\mu_{k,on} - \mu_{k,off} = -1.06 +$

0.71 = 0.35, and also a high value of $\rho_k \frac{\sigma_{k,off}}{\sigma_{k,on}} = 1$. Although a big discrepancy exists between online and offline partworths, given a respondent's online responses, and given good estimates of $\mu_{k,on}$ and $\mu_{k,off}$, we can predict the individual-level $\beta_{k,off}$ well.

5.3 Relative information gain in offline and online data

Section 4 shows that our proposed approach, which combines offline and online conjoint results, yields more accurate individual- and aggregate-level purchase predictions than does pure online-only conjoint. As our objective is to obtain accurate predictions of offline purchase behavior, an alternate approach is to forgo the online task altogether, and simply conduct offline conjoint studies. By requiring the respondents to evaluate physical prototypes, these studies are closer to the real-world purchasing context and therefore preferable. The tradeoff, of course, is that conducting a conjoint study offline is significantly costlier per participant.

In this section, using the unrestricted linear mixed model described in Equations (3) and (4), we compare our proposed approach to conducting an offline-only conjoint on the *precisions* of the offline parameter estimates they obtain. We measure the precision of a parameter estimate in terms of the asymptotic variance of the corresponding maximum-likelihood (ML) estimator, obtained from the inverse of the Fisher information matrix, to be described shortly. Lower variance values indicate higher precisions. To conduct the comparison on equal footings, we focus on settings in which the costs of the two approaches are equalized; of course, without a cost constraint, conducting an offline-only conjoint should yield better performance.

The key insight from our analysis is that our proposed approach can take advantage of the cost differential between conducting online and offline conjoint studies to obtain more precise estimates than those yielded by an offline-only conjoint. When there is a cost differential, our approach uses a small share of the budget to collect a large sample of “noisy” (online) data as opposed to a small sample of “cleaner” (offline) data. The larger sample size more than makes up for the noise in the data to yield more precise estimates, for the values of cost differentials observed in practice.

Setup. We considered the following setup for our analysis: Suppose the cost to the firm of a single online respondent is c_{on}, and the cost of a single offline respondent is c_{off}, such that $c_{off} > c_{on}$. Let $Q = c_{off}/c_{on}$ denote the cost multiplier of an offline respondent with respect to an online one. Given cost values to be discussed shortly, we compared two study designs: a ***combined*** design and an ***offline-only*** design. The combined design collects offline and online data for a set of respondents, and online-only data for an additional set of respondents. Thus, all respondents complete the online study, and we let N_{on} represent the size of this set. Offline data are collected for N_{off} respondents. The offline-only design collects offline data only for N respondents. In order to equalize the costs of the two designs, the following relationship should be satisfied:

(19) $$c_{off}N = c_{on}N_{on} + c_{off}N_{off}$$

This implies that:

(20) $$N = N_{off} + \Delta, \text{ where } \Delta = N_{on}/Q$$

In other words, we can collect offline-only data, at the same total cost, for Δ more respondents than in the combined design.

To obtain a realistic estimate to the multiplier Q, we obtained price quotes from market research firms for a commercial offline conjoint study. The costs involve: payment to participants, the hourly rate of an experimenter (including salary, benefits, and overhead), and recruiting costs. The total comes to \$100-\$150 per participant. Online participants can be obtained for \$2-\$3 per participant. With these cost values, we set N_{on} = 122, N_{off} = 61 and varied the cost multiplier Q to take values from the set {50, 30, 15}. While the value of Q = 50 is reasonable given the quotes we obtained, we also considered the value Q = 15 as a lower-end estimate, to capture any settings in which conducting an offline conjoint is relatively cheap. Lastly we also considered an intermediate multiplier of 30. With these values of Q, we obtained $\Delta = N_{on}/Q$ to be approximately 2, 4, and 8 (corresponding to the values of Q= 50, 30, and 15). Thus, while we kept the benchmark case at 61 online and offline, plus 61 online respondents, we compared this benchmark to 63, 65, and 69 offline respondents only. For example, for Q = 50, given the figures of \$2 and \$100 for online and offline cost per respondent, the total budget remains the same for the two studies: approximately \$6,300 (see Equation 19).

Results and discussion. Table 6 shows the asymptotic variances corresponding to the ML estimates of the offline partworths, as the cost multiplier is varied over Q = 50, 30, 15. We obtained these values by taking the inverses of the Fisher information matrices computed for the two designs; the details of the computations are presented in Appendix D.

Table 6: Asymptotic variance performance

	Combined design	**Offline-only design**		
	N_{on} = 122 N_{off} = 61	N = 63 (Q = 50)	N = 65 (Q = 30)	N = 69 (Q = 15)
reflective	0.0161	0.0160	0.0155	0.0146
colorful	0.0147	**0.0203**	**0.0196**	**0.0185**
blue	0.0062	**0.0064**	**0.0062**	0.0059
size	0.0059	**0.0067**	**0.0065**	**0.0061**
price	1.27E-06	**1.33E-06**	**1.29E-06**	1.22E-6

strap pad	0.0038	**0.0042**	**0.0040**	0.0038
water bottle	0.00249	**0.00254**	0.0025	0.0023
divider	0.00379	0.00378	0.0037	0.0034
laptop	0.0072	**0.0076**	**0.0074**	0.0070

We note that when cost multiplier Q = 50 (i.e., conducting an offline conjoint is relatively costly, as market prices indicate), our proposed combined design yields more precise estimates (i.e., with lower asymptotic variances) of the offline partworths than does the offline-only design for *all* but "reflective" and "divider" features. The reduction in the variance due to the combined design can be significant: as much as 28% for the partworth "colorful". The increase in the variance, on the other hand, is less than 1% for "reflective" and "divider" features. These findings enable us to conclude that when conducting an offline conjoint is significantly more costly than conducting an online one, using some portion of the budget to conduct an online conjoint for a large sample of respondents can provide more precise estimates of the offline partworths. In these settings, the loss in the "quality" of the data per respondent from collecting online instead of offline data is more than made up for by the ability to collect a much larger sample.

When the cost multiplier Q reduces to 30, the combined design results in lower variances for six features, with a reduction of up to 24% for the feature "colorful", and higher variances for the remaining three features, with a maximum increase of less than 4%. Therefore, even if the cost of conducting an offline conjoint dropped by 40%, the combined design offers better precision overall when compared to the offline-only design. It is only when conducting an offline conjoint becomes significantly cheaper, i.e., when Q = 15, that the offline-only conjoint outperforms the combined design. This analysis shows how the benefits of the combined approach over the offline-only conjoint hinge on the cost differential Q. In most practical settings, we expect *Q* to be reasonably large, given the ease of finding respondents online through crowdsourcing platforms, such as MTurk. As a result, the comparison corresponding to *Q* = 50 is more indicative of what we expect to see in practice.

6. Conclusions and implications

In this work, we challenged the common implicit assumption in preference elicitation that findings from online studies can accurately predict offline purchase behavior. We compared consumers' product evaluations in an online conjoint study with verbal product descriptions and pictures with those of the same consumers in an offline study with physical products. We found that the majority of partworth parameters changed significantly between online and offline studies. This discrepancy will lead models trained on data from online studies only, to have diminished predictive ability for offline behavior. We recognize, however, that conducting online preference elicitation is significantly cheaper than conducting offline preference elicitation. Therefore, we proposed and tested a hybrid solution: supplementing an online conjoint study with a small set of participants who complete both online and offline preference elicitation. We tested two data-fusion techniques that use the data from an online study completed by a "large" number of respondents, supplemented by an offline study completed by a "small" subset of the respondents. The techniques predict a respondent's offline preference partworths when given her online partworths. In the empirical application, we demonstrated that our data-fusion techniques result in more accurate predictions of respondents' offline choices.

Our study consisted of two conditions in which participants completed the online and offline conjoint tasks in differing orders, allowing us to gain further insight into the source of the discrepancy. The results suggest that two key factors cause respondents to behave differently when evaluating products online versus offline: information gained by physically and visually examining the product, and differing relative attribute salience in the two formats. Note that neither of these factors is related to consumer preferences actually differing between online and offline settings. What the conjoint partworths more precisely represent are decision rules, or the extent to which a product's having a certain feature increases the probability of its being chosen.

The decision rule may not perfectly capture a consumer's preference. For example, making a certain feature, such as strap pad, more salient, such as by increasing the font size of that feature in the product description, will increase the weight that the strap pad carries in the respondent's decision. Clearly, the larger font does not increase the respondent's actual preference for the feature: S/he does not start to like the strap pad any more or less than before. It does, however, increase the role it plays in the consumer decision, which is what the partworths capture.

One of the limitations of our approach is that it is applicable to the cases in which a prototype is available for preference-elicitation techniques. It thus does not apply to services, but rather to physical goods purchased in brick-and-mortar stores only. Clearly, for a service such as cellular, the packages (usually three-tier assortments) do not lend themselves to information gain offline as compared to online, and attribute salience should not exist, as both are presented in list form. Neither can our approach be used to forecast demand for radical innovations, for which physical prototypes are not yet available. Note that the discrepancy between online and offline attribute evaluations might be consumer specific, and thus a possible future avenue of research could examine consumer characteristics, and in particular level of familiarity with the product category, that might help explain this discrepancy.

In this paper, we used primary data to carefully control for all factors, and to home in on the online/offline distinction. But the higher-level problem of predicting a consumer's offline preferences, given the same consumer's online preferences, and *other* consumers' online and offline preferences, has implications beyond online preference elicitation. Typically, the firm has, or can obtain, some data on both online and offline preferences for customers who have a history with both, as is depicted in Figure 3.

Figure 3: Schematic data available to a typical online retailer

	Existing customers	New customers
Online	data used for estimation	
Offline		prediction task

These preferences could be estimated from secondary sources, such as past purchases, clicks, or returns data. Consider mixed retailers who sell both online and offline, such as Warby Parker, Zappos, or Bonobos, or online-only retailers: Both are affected by the discrepancy of online and offline product evaluation due to "showrooming" and the prevalence of flexible return policies. For instance, when consumers purchase from online/mixed retailers, they may decide what to order based on their online evaluation of the available items. However, once they receive their order, consumers determine what they want to keep and what to return based on physical evaluation. Because of the generous return policies offered by many retailers, customers may try on several items before purchasing one. To apply our methods, an online/mixed retailer can use the online and offline preference data obtained from the items that a given customer ordered online, and the items that s/he decided to return after physical evaluation. Such customers can be considered the training set, as they provide sufficient data to calibrate the discrepancy between online and offline partworths. For a new customer, who has not yet evaluated the firm's products physically, the firm may have data on online preferences only. In this case, the firm can apply an approach similar to ours to predict what the customer will prefer upon physical examination of the products. With this prediction, the firm can better manage returns, or may recommend products that the customer is likely to prefer in person. For use of a similar approach, see Dzyabura, El Kihal, and Ibragimov (2017).

References

Abu-Mostafa, Yaser S., Malik Magdon-Ismail, and Hsuan-Tien Lin. 2012. *Learning from data*. Vol. 4. New York: AMLBook.

Allison, Paul D. and Nicholas A. Christakis. 1994. Logit models for sets of ranked items. *Sociological Methodology*, 24, 199-228.

Beggs, Steven; Scott Cardell, and Jerry Hausman. 1981. Assessing the potential demand for electric cars. *Journal of Econometrics*, 17(1), 1-19.

Bell, Robert M. and Yehuda Koren. 2007. Scalable collaborative filtering with jointly derived neighborhood interpolation weights. *IEEE International Conference on Data Mining* (ICDM07), 43-52.

Ben-Akiva, M., M. Bradley, T. Morikawa, J. Benjamin, T. Novak, H. Oppewal, and V. Rao. 1994. Combining revealed and stated preferences data. *Marketing Letters*, 5(4), 335-349.

Bettman, James R., Mary Frances Luce, and John W. Payne. 1998. Constructive consumer choice processes, *Journal of Consumer Research*, 25(3), 187-217.

Bhat, Chandra R. and Saul Castelar. 2002. A unified mixed logit framework for modeling revealed and stated preferences: Formulation and application to congestion pricing analysis in the San Francisco Bay Area. *Transportation Research Part B: Methodological*, 36(7), 593-616.

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

Chapman, Randall G. and Richard Staelin. 1982. Exploiting rank ordered choice set data within the stochastic utility model. *Journal of Marketing Research*, 29(3), 288-301.

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

Ding, Min, John R. Hauser, Songting Dong, Daria Dzyabura, Zhilin Yang, Chenting Su, and Steven P. Gaskin. 2011. Unstructured direct elicitation of decision rules. *Journal of Marketing Research*, 48(1), 116-127.

Dzyabura, Daria, Siham El Kihal, and Marat Ibragimov. 2017. Product return management in omnichannel retail: Leveraging the predictive ability of product images. Working paper.

Dzyabura, Daria and Srikanth Jagabathula. 2017. Offline assortment optimization in the presence of an online channel. *Management Science*, forthcoming.

Dzyabura, Daria and John R. Hauser. 2011. Active machine learning for consideration heuristics. *Marketing Science*, 30(5), 801-819.

Feinberg, Fred M., Thomas Kinnear, and James R. Taylor. *Modern marketing research: Concepts, methods, and cases*. Cengage Learning, 2012.

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

Gilbride, Timothy J., Peter J. Lenk, and Jeff D. Brazell. 2008. Market share constraints and the loss function in choice-based conjoint analysis. *Marketing Science*, 27(6), 995-1011.

Green, Paul E. and Vithala R. Rao. 1971. Conjoint measurement for quantifying judgmental data. *Journal of Marketing Research*, 8(3), 355-363.

Hausman, Jerry A. and Paul A. Ruud. 1987. Specifying and testing econometric models for rank-ordered data. *Journal of Econometrics*, 34(1), 83-104.

Hauser, John R., Olivier Toubia, Theodoros Evgeniou, Rene Befurt, and Daria Dzyabura. 2010. Disjunctions of conjunctions, cognitive simplicity, and consideration sets. *Journal of Marketing Research*, 47(3), 485-496.

Higgins, E. Tory. 1996. Knowledge activation: Accessibility, applicability, and salience. In E. T. Higgins & A. W. Kruglanski (eds.), *Social psychology: Handbook of basic principles* (133-168). New York: Guilford Press.

Huber, Joel and Klaus Zwerina. 1996. The importance of utility balance in efficient choice designs. *Journal of Marketing Research*, 33(3), 307-317.

Kohli, Rajeev and Sukumar, Ramamirtham. 1990. Heuristics for product-line design using conjoint analysis. *Management Science*, 36(12), 1464-1478.

Kök, A. Gürhan, Marshall L. Fisher, and Ramnath Vaidyanathan. 2015. Assortment planning: Review of literature and industry practice. Pp. 175-236, *Retail supply chain management*. Springer.

Kuhfeld, Warren F., Randall D. Tobias, and Mark Garratt. 1994. Efficient experimental design with marketing research applications. *Journal of Marketing Research*, 31(4), 545-557.

Lenk, Peter J., Wayne S. DeSarbo, Paul E. Green, and Martin R. Young. 1996. Hierarchical Bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs. *Marketing Science*, 15(2), 173-191.

Luo, Lan, P. K. Kannan, and Brian T. Ratchford. 2008. Incorporating subjective characteristics in product design and evaluations. *Journal of Marketing Research*, 45(2), 182-194.

Netzer, Oded, Olivier Toubia, Eric T. Bradlow, Ely Dahan, Theodoros Evgeniou, Fred M. Feinberg, and Eleanor M. Feit. 2008. Beyond conjoint analysis: Advances in preference measurement. *Marketing Letters*, 19(3-4), 337-354.

Orme, Bryan and Rich Johnson. 2006. External effect adjustments in conjoint analysis. *SAWTOOTH SOFTWARE CONFERENCE*, Delray Beach, FLA.

Peck, Joann and Terry L. Childers. 2003. To have and to hold: The influence of haptic information on product judgments. *Journal of Marketing* 67(2), 35-48.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, and Mathieu Blondel et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 2825-2830.

PricewaterhouseCoopers LLP. (2015). *Physical Store Beats Online as Preferred Purchase Destination for U.S. Shoppers, According to PwC* [press release]. Retrieved from http://www.prnewswire.com/news-releases/physical-store-beats-online-as-preferred-purchase-destination-for-us-shoppers-according-to-pwc-300032566.html

Punj, Girish N. and Richard Staelin. 1978. The choice process for graduate business schools. *Journal of Marketing Research*, 15(4), 588-598.

Rossi, Peter E., Greg M. Allenby, and Rob McCulloch. 2012. *Bayesian statistics and marketing*. John Wiley & Sons.

She, Jinjuan and Erin F. MacDonald. 2013. Trigger features on prototypes increase preference for sustainability. *ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Portland, OR. American Society of Mechanical Engineers V005T06A043-V005T06A054.

Srinivasan, V., William S. Lovejoy, and David Beach. 1997. Integrated product design for marketability and manufacturing. *Journal of Marketing Research*, 34(1), 154-163.

Swait, Joffre and Rick L. Andrews. 2003. Enriching scanner panel models with choice experiments. *Marketing Science*, 22(4), 442-460.

Wu, Xindong, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, and Geoffrey J. McLachlan et al. Top 10 algorithms in data mining. *Knowledge and Information Systems* 14, no. 1 (2008): 1-37.

US Census Bureau News: Quarterly retail e-commerce sales, 2nd quarter 2017. https://www.census.gov/retail/mrts/www/data/pdf/ec_current.pdf

Appendix A: Linear Mixed-Effects Model – Covariance Parameters

Variance and covariance parameters of the unrestricted model in Equation 1, estimated on 122 participants in Condition 1 (online first).

Attribute	Level	Online Variance ($\sigma^2_{k,on}$)	Offline Variance ($\sigma^2_{k,off}$)	Online Offline Covariance ($\sigma_{k,on,off}$)
Exterior design	Reflective	0.27	0.72	0.16
	Colorful	0.90	1.03	0.90
	Blue	0.15	0.16	0.14
	Black			
Size	Large	0.16	0.31	0.18
	Small			
Price	$120, $140, $160, $180	2.28E-5	2.39E-5	2.29E-5
Strap pad	Yes	0.15	0.14	0.12
	No			
Water bottle pocket	Yes	0.06	0.05	0.05
	No			
Interior compartments	Divider for files	0.06	0.05	0.05
	Crater laptop sleeve	0.24	0.31	0.17
	Empty bucket/no dividers			

Appendix B: ICL correction with restricted covariance matrix

The ICL method computes the expected offline ratings conditioned on all the observed data. We exploit the properties of multivariate normal distributions in order to compute the conditional expectations in closed form. Specifically, recall that we assume that participant *i* samples the online and offline partworths, $\beta_{i,k,on}$ and $\beta_{i,k,off}$, of feature k jointly from a bivariate normal distribution:

$$\boldsymbol{\beta_{i,k}} \sim \mathcal{N}(\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})$$

$$\boldsymbol{\beta_{i,k}} = \begin{bmatrix} \beta_{i,k,on} \\ \beta_{i,k,off} \end{bmatrix}, \boldsymbol{\mu_k} = \begin{bmatrix} \mu_{k,on} \\ \mu_{k,off} \end{bmatrix}, \boldsymbol{\Sigma_k} = \begin{bmatrix} \sigma_{k,on}^2 & \sigma_{k,on,off} \\ \sigma_{k,on,off} & \sigma_{k,off}^2 \end{bmatrix},$$

where $\sigma_{k,on,off}$ is the covariance between the online and offline partworths of feature *k*. The assumption embedded in the covariance matrix is that there are no correlations among various features, that is, we fix at zero the elements of $\boldsymbol{\Sigma}$ that correspond to $cov(\beta_{i,k,t}, \beta_{i,k',t'})$ for $k \neq k'$, and for all t and t'.

We use observed data to determine the maximum likelihood estimates of the population-level parameters $\mu_{k,off}$, $\mu_{k,on}$, $\sigma_{k,off}, \sigma_{k,on}$, and $\sigma_{k,on,off}$ for each feature *k*. Note that the data from the group of respondents who completed both the online and offline tasks enables us to estimate the covariance parameters. Given the population-level parameters, we can show that the conditional distribution of $\beta_{i,k,off}$ given $\beta_{i,kj,on}$ is a normal one, with mean $\mu_{i,k,off|on}$ and variance $\sigma_{i,k,off|on}^2$ as given by

$$\mu_{i,k,off|on} = \mu_{k,off} + \rho_k \frac{\sigma_{k,off}}{\sigma_{k,on}} (\beta_{i,k,on} - \mu_{k,on})$$

$$\sigma_{i,k,off|on} = \sigma_{k,off} \sqrt{1 - \rho_k^2}.$$

where ρ_k is the correlation coefficient between feature *k*'s online and offline partworths.

Note that $\mu_{i,k,off|on}$ is also the maximum likelihood estimator of $\beta_{i,k,off}$ conditioned on $\beta_{i,k,on}$ due to normality. Therefore, under this model, conditioned on $\beta_{i,k,on}$, the maximum likelihood estimates of a respondent's offline partworths are given by:

$$\hat{\beta}_{i,k,off} = \mu_{k,off} + \rho_k \frac{\sigma_{k,off}}{\sigma_{k,on}} (\beta_{i,k,on} - \mu_{k,on}).$$

Appendix C: HB estimation

Here we specify the details of the hierarchical Bayes estimation of the online and offline data: the prior covariance, and the posterior estimates of the partworths and the covariance.

We coded the online and offline attributes to be levels of a single attribute. For example, the online and offline color partworths to be 8 levels of one attribute, online and offline partworths of interior compartments to be 6 levels of one attribute, etc. The reason for making them different levels of the same attribute rather than two different attributes (e.g. “online color” with 4 levels and “offline color” with 4 levels) is that the estimation requires that all products have one level of each attribute.

Table C1: Prior covariance*

Color	1.75	-0.2	-0.2	-0.2	0.3	-0.2	-0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	-0.2	1.75	-0.25	-0.2	-0.2	0.3	-0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	-0.2	-0.25	1.75	-0.2	-0.25	-0.2	0.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	-0.2	-0.2	-0.2	1.75	-0.2	-0.2	-0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0.3	-0.2	-0.25	-0.2	1.75	-0.2	-0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	-0.2	0.3	-0.2	-0.2	-0.2	1.75	-0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	-0.2	-0.2	0.3	-0.2	-0.2	-0.2	1.75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Size	0	0	0	0	0	0	0	1.5	-0.5	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	-0.5	1.5	-0.5	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0.5	-0.5	1.5	0	0	0	0	0	0	0	0	0	0	0	0	0
Price	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0
Strap Pad	0	0	0	0	0	0	0	0	0	0	0	0	1.5	-0.5	0.5	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	-0.5	1.5	-0.5	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0.5	-0.5	1.5	0	0	0	0	0	0	0	0
Water bottle pocket	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.5	-0.5	0.5	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.5	1.5	-0.5	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	-0.5	1.5	0	0	0	0	0
Interior compartments	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.67	-0.33	-0.33	0.33	-0.33
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.33	1.67	-0.33	-0.33	0.33
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.33	-0.33	1.67	-0.33	-0.33
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.33	-0.33	-0.33	1.67	-0.33
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.33	0.33	-0.33	-0.33	1.67

* The levels of each attribute are given in Table C1.

Table C2: Posterior covariance estimate*

Color	1.92	-0.50	-0.75	0.25	0.67	0.02	-1.35	0.09	-0.04	-0.03	0.00	0.00	0.09	0.27	-0.06	-0.06	0.06	0.09	-0.04	0.21	0.25	-0.08	-0.15
	-0.50	2.29	-0.41	-0.57	-0.44	0.48	-0.03	-0.04	0.03	-0.02	0.02	0.00	-0.09	-0.13	0.13	0.07	0.01	-0.01	0.19	-0.08	0.08	0.01	0.00
	-0.75	-0.41	3.65	-0.64	-1.35	-1.21	2.36	-0.01	0.18	0.02	-0.04	0.00	-0.01	-0.13	-0.16	-0.20	-0.28	-0.04	-0.28	-0.19	-0.24	0.26	0.22
	0.25	-0.57	-0.64	1.89	0.14	-0.20	-1.06	0.11	0.01	-0.03	0.02	0.00	0.15	0.12	-0.03	0.09	0.09	-0.01	0.17	0.03	-0.04	0.00	-0.11
	0.67	-0.44	-1.35	0.14	2.20	0.18	-1.63	0.03	-0.18	-0.09	0.00	0.00	0.07	0.14	0.01	0.00	0.11	0.04	-0.01	0.13	0.17	-0.21	-0.14
	0.02	0.48	-1.21	-0.20	0.18	2.10	-1.12	-0.10	-0.07	0.05	0.03	0.00	-0.04	0.02	0.08	0.17	0.13	-0.01	0.28	0.06	0.11	-0.08	-0.11
	-1.35	-0.03	2.36	-1.06	-1.63	-1.12	4.77	-0.13	0.29	0.10	-0.02	0.01	-0.18	-0.44	0.06	-0.14	-0.32	0.01	-0.29	-0.26	-0.47	0.46	0.41
Size	0.09	-0.04	-0.01	0.11	0.03	-0.10	-0.13	1.29	-0.41	0.24	-0.02	0.00	0.10	-0.07	0.07	-0.01	0.03	0.00	-0.15	0.02	-0.05	0.02	0.12
	-0.04	0.03	0.18	0.01	-0.18	-0.07	0.29	-0.41	1.33	-0.44	0.00	0.00	-0.01	0.01	0.04	0.01	0.04	-0.05	0.07	-0.15	-0.03	0.09	0.00
	-0.03	-0.02	0.02	-0.03	-0.09	0.05	0.10	0.24	-0.44	1.14	0.01	0.00	-0.01	-0.06	0.11	0.07	-0.02	0.01	0.06	0.04	-0.03	0.01	0.01
Price	0.00	0.02	-0.04	0.02	0.00	0.03	-0.02	-0.02	0.00	0.01	0.90	0.36	0.00	-0.03	0.03	0.04	0.00	0.03	0.00	-0.01	0.04	0.01	-0.02
	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.36	0.72	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
Strap pad	0.09	-0.09	-0.01	0.15	0.07	-0.04	-0.18	0.10	-0.01	-0.01	0.00	0.00	1.09	-0.19	0.22	0.05	-0.03	0.09	0.09	0.00	0.04	0.05	-0.04
	0.27	-0.13	-0.13	0.12	0.14	0.02	-0.44	-0.07	0.01	-0.06	-0.03	0.00	-0.19	1.26	-0.42	-0.10	0.05	0.02	0.02	0.10	0.08	-0.10	-0.13
	-0.06	0.13	-0.16	-0.03	0.01	0.08	0.06	0.07	0.04	0.11	0.03	0.00	0.22	-0.42	1.08	0.10	0.05	0.07	0.01	-0.05	0.02	0.10	0.00
Water bottle pocket	-0.06	0.07	-0.20	0.09	0.00	0.17	-0.14	-0.01	0.01	0.07	0.04	0.00	0.05	-0.10	0.10	1.07	-0.11	0.08	0.15	-0.02	-0.05	-0.01	-0.01
	0.06	0.01	-0.28	0.09	0.11	0.13	-0.32	0.03	0.04	-0.02	0.00	0.00	-0.03	0.05	0.05	-0.11	1.08	-0.43	-0.02	0.05	0.04	-0.12	0.06
	0.09	-0.01	-0.04	-0.01	0.04	-0.01	0.01	0.00	-0.05	0.01	0.03	0.00	0.09	0.02	0.07	0.08	-0.43	1.09	0.06	-0.02	0.06	0.13	-0.14
Interior compartment	-0.04	0.19	-0.28	0.17	-0.01	0.28	-0.29	-0.15	0.07	0.06	0.00	0.00	0.09	0.02	0.01	0.15	-0.02	0.06	1.53	-0.23	-0.31	0.21	-0.30
	0.21	-0.08	-0.19	0.03	0.13	0.06	-0.26	0.02	-0.15	0.04	-0.01	0.00	0.00	0.10	-0.05	-0.02	0.05	-0.02	-0.23	1.21	-0.01	-0.37	0.08
	0.25	0.08	-0.24	-0.04	0.17	0.11	-0.47	-0.05	-0.03	-0.03	0.04	0.00	0.04	0.08	0.02	-0.05	0.04	0.06	-0.31	-0.01	1.45	-0.37	-0.45
	-0.08	0.01	0.26	0.00	-0.21	-0.08	0.46	0.02	0.09	0.01	0.01	0.01	0.05	-0.10	0.10	-0.01	-0.12	0.13	0.21	-0.37	-0.37	1.50	-0.19
	-0.15	0.00	0.22	-0.11	-0.14	-0.11	0.41	0.12	0.00	0.01	-0.02	0.00	-0.04	-0.13	0.00	-0.01	0.06	-0.14	-0.30	0.08	-0.45	-0.19	1.30

* The level of each attribute is given in Table C1.

Appendix D: Computation of the Asymptotic Variances

We now present the details of the computations we carried out to obtain the asymptotic variances corresponding to the maximum likelihood estimates of the offline partworths under the linear mixed effects (LME) model. We compute the variances by inverting the Fisher Information (FI) matrix, which we compute following Theorem 1 in Lenk et al. 1996.

Our computations were carried out on the data collected as part of the study, described in Section 3. The study was carried out on $n = 20$ products, selected using the D-optimal study-design criterion. Product j is described by the length K feature vector $\boldsymbol{x}_j$, obtained by dummy-coding 13 discrete attribute levels into $K = 9$ features. Collecting the feature vectors together, we obtain the following design matrix:

$$\boldsymbol{X} = \begin{bmatrix} — & (\boldsymbol{x}_1)^\top & — \\ — & (\boldsymbol{x}_2)^\top & — \\ & \vdots & \\ — & (\boldsymbol{x}_n)^\top & — \end{bmatrix},$$

In our study, each respondent is exposed to the same design $\boldsymbol{X}$ and asked to rate the products in an offline, online, or an online followed by offline conjoint. We assume that the rating assigned by respondent i for product j follows the following model:

$$u_{i,j,\text{off}} = \beta_{i,0} + \delta_i + \sum_{k=1}^{K} \beta_{i,k,\text{off}}\, x_{jk} + \varepsilon_{i,j,\text{off}}, \text{ if } j \text{ was evaluated offline, and}$$

$$u_{i,j,\text{on}} = \beta_{i,0} + \sum_{k=1}^{K} \beta_{i,k,\text{on}}\, x_{jk} + \varepsilon_{i,j,\text{on}}, \text{ if } j \text{ was evaluated online,}$$

where $\beta_{i,0}$ is the intercept term, δ_i is the offline fixed effect, and $\boldsymbol{\beta}_{i,\text{off}} = [\beta_{i,1,\text{off}}, \ldots, \beta_{i,K,\text{off}}]^\top$ and $\boldsymbol{\beta}_{i,\text{on}} = [\beta_{i,1,\text{on}}, \ldots, \beta_{i,K,\text{on}}]^\top$ are the offline and online partworth vectors, respectively. We assume that individual i independently samples $\beta_{i,0}$ according to $N(\mu_0, \sigma_0^2)$, and δ_i according to $N(\mu_f, \sigma_f^2)$, $\varepsilon_{i,j,\text{on}}$ and $\varepsilon_{i,j,\text{off}}$ according to $N(0, \sigma^2)$, for all j, and $\boldsymbol{\beta}_i = [\boldsymbol{\beta}_{i,\text{off}}^\top, \boldsymbol{\beta}_{i,\text{on}}^\top]$ according to $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = [\boldsymbol{\mu}_{\text{off}}^\top, \boldsymbol{\mu}_{\text{on}}^\top]^\top$ and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{\text{off}} & \boldsymbol{\Sigma}_{\text{on,off}} \\ \boldsymbol{\Sigma}_{\text{on,off}} & \boldsymbol{\Sigma}_{\text{on}} \end{bmatrix} \text{ with}$$

$$\boldsymbol{\Sigma}_{\text{off}} = \text{diag}(\sigma_{1,\text{off}}^2, \ldots, \sigma_{K,\text{off}}^2), \boldsymbol{\Sigma}_{\text{on}} = \text{diag}(\sigma_{1,\text{on}}^2, \ldots, \sigma_{K,\text{on}}^2), \text{ and } \boldsymbol{\Sigma}_{\text{on,off}} = \text{diag}(\sigma_{1,\text{on,off}}, \ldots, \sigma_{K,\text{on,off}}),$$

where $\text{diag}(\boldsymbol{v})$ denotes the diagonal matrix with $\boldsymbol{v}$ as the diagonal.

When a respondent evaluates the n products in an offline conjoint and then the same n products in an online conjoint, the relation between the ratings collected and the underlying model parameters can be written more compactly as

$$\boldsymbol{u}_i = \begin{bmatrix} \boldsymbol{u}_{i,\text{off}} \\ \boldsymbol{u}_{i,\text{on}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_{n\times 1} & \mathbf{1}_{n\times 1} & \boldsymbol{X} & \mathbf{0}_{n\times K} \\ \mathbf{1}_{n\times 1} & \mathbf{0}_{n\times 1} & \mathbf{0}_{n\times K} & \boldsymbol{X} \end{bmatrix} \begin{bmatrix} \beta_{i,0} \\ \delta_i \\ \boldsymbol{\beta}_{i,\text{off}} \\ \boldsymbol{\beta}_{i,\text{on}} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_{i,\text{off}} \\ \boldsymbol{\varepsilon}_{i,\text{on}} \end{bmatrix},$$

where $\boldsymbol{u}_{i,t} = [u_{i,1,t}, \dots, u_{i,K,t}]^\top$, $\boldsymbol{\varepsilon}_{i,t} = [\varepsilon_{i,1,t}, \dots, \varepsilon_{i,K,t}]^\top$, for $t \in \{\text{off}, \text{on}\}$. For compactness of notation, we define

$$\boldsymbol{X}_{\text{full}} = \begin{bmatrix} \mathbf{1}_{n\times 1} & \mathbf{1}_{n\times 1} & \boldsymbol{X} & \mathbf{0}_{n\times K} \\ \mathbf{1}_{n\times 1} & \mathbf{0}_{n\times 1} & \mathbf{0}_{n\times K} & \boldsymbol{X} \end{bmatrix}$$

and say that the respondent evaluated the design $\boldsymbol{X}_{\text{full}}$ when s/he completes an online followed by an offline conjoint. Similarly, we define

$$\boldsymbol{X}_{\text{off}} = \begin{bmatrix} \mathbf{1}_{n\times 1} & \mathbf{1}_{n\times 1} & \boldsymbol{X} & \mathbf{0}_{n\times K} \\ \mathbf{0}_{n\times 1} & \mathbf{0}_{n\times 1} & \mathbf{0}_{n\times K} & \mathbf{0}_{n\times K} \end{bmatrix} \text{ and } \boldsymbol{X}_{\text{on}} = \begin{bmatrix} \mathbf{0}_{n\times 1} & \mathbf{0}_{n\times 1} & \mathbf{0}_{n\times K} & \mathbf{0}_{n\times K} \\ \mathbf{1}_{n\times 1} & \mathbf{0}_{n\times 1} & \mathbf{0}_{n\times K} & \boldsymbol{X} \end{bmatrix},$$

where $\boldsymbol{X}_{\text{off}}$ (respectively, $\boldsymbol{X}_{\text{on}}$) is obtained by replacing the upper (respectively, lower) block row with all zeros. We say that a respondent evaluated the design $\boldsymbol{X}_{\text{off}}$ (respectively, $\boldsymbol{X}_{\text{on}}$) if s/he completes only an offline (respectively, online) conjoint. Finally, define

$$\boldsymbol{\Sigma}_{\text{full}} = \begin{bmatrix} \sigma_0^2 & 0 & \mathbf{0}_{1\times 2K} \\ 0 & \sigma_f^2 & \mathbf{0}_{1\times 2K} \\ \mathbf{0}_{2K\times 1} & \mathbf{0}_{2K\times 1} & \boldsymbol{\Sigma} \end{bmatrix}.$$

We now invoke Theorem 1 from Lenk et al. 1996 to compute the FI matrix corresponding to the paramters $\tilde{\mu} = [\mu_0, \mu_f, \boldsymbol{\mu}^\top]^\top$. Under the combined design, N_{off} respondents complete an online followed by offline conjoint and N_{on} respondents complete only the online conjoint. It can be shown that the FI under this design is given by

$$\text{FI}_{\text{full}} = N_{\text{off}} \cdot \boldsymbol{X}_{\text{full}}^\top \Lambda_{\text{full}}^{-1} \boldsymbol{X}_{\text{full}} + N_{\text{on}} \cdot \boldsymbol{X}_{\text{on}}^\top \Lambda_{\text{on}}^{-1} \boldsymbol{X}_{\text{on}},$$

where

$$\boldsymbol{\Lambda}_{\text{full}} = \sigma^2 \boldsymbol{I}_{2n\times 2n} + \boldsymbol{X}_{\text{full}} \Sigma_{\text{full}} \boldsymbol{X}_{\text{full}}^\top \text{ and } \boldsymbol{\Lambda}_{\text{on}} = \sigma^2 \boldsymbol{I}_{2n\times 2n} + \boldsymbol{X}_{\text{on}} \Sigma_{\text{full}} \boldsymbol{X}_{\text{on}}^\top$$

with $\boldsymbol{I}_{m\times m}$ is an $m \times m$ identity matrix for any m. The first term in FI_{full} corresponds to the FI from the respondents who complete an online followed by an offline conjoint, equivalently, a conjoint on the design $\boldsymbol{X}_{\text{full}}$. The second term, on the other hand, corresponds to the FI from the respondents who complete only an online conjoint, equivalently, a conjoint on the design $\boldsymbol{X}_{\text{on}}$. The counts N_{off} and N_{on} factor out because each of the N_{off} (respectively, N_{on}) respondents evaluate the same set of profiles $\boldsymbol{X}_{\text{full}}$ (respectively, $\boldsymbol{X}_{\text{on}}$).

In a similar manner, the FI under the offline only conjoint can be shown to be given by

$$\text{FI}_{\text{off}} = N \cdot \boldsymbol{X}_{\text{off}}^\top \Lambda_{\text{off}}^{-1} \boldsymbol{X}_{\text{off}},$$

where

$$\boldsymbol{\Lambda}_{\text{off}} = \sigma^2 \boldsymbol{I}_{2n\times 2n} + \boldsymbol{X}_{\text{off}} \Sigma_{\text{full}} \boldsymbol{X}_{\text{off}}^\top$$

and N is the number of respondents who complete an offline conjoint on the set of profiles $\boldsymbol{X}_{\text{off}}$.

For computing the above FI matrices, we used design matrix $\boldsymbol{X}$ reported in Section 3 and $\boldsymbol{\Sigma}_{\text{full}}$ reported in Appendix A. Once we compute the FI matrices, we obtained the asymptotic variances under the combined and offline only designs corresponding to the estimate of $\mu_{k,\text{off}}$ by taking the diagonal element corresponding to $\mu_{k,\text{off}}$ in $\text{FI}_{\text{full}}^{-1}$ and $\text{FI}_{\text{off}}^{-1}$ respectively.