Online Appendix:
Scalable Bundle Recommendations: A Large-Scale Field Experiment

A Embeddings Model

A.1 Illustrative Example of Product Embeddings Computation

To illustrate the mechanics of our product embeddings model, we present a simplified example using 8 items. This example explains the key steps of the embedding process: initializing matrices, iterating through purchase baskets, updating embeddings, and using the final embeddings to identify potential complements/substitutes. We only focus on purchase baskets, and hence purchase embeddings, in this example. The training and update scheme for consideration set embeddings is the same.

Step 1: Observed Purchase Baskets

Consider the following 8 grocery products:

1. Coffee
2. Cream
3. Sugar
4. Bread
5. Butter
6. Jam
7. Cereal
8. Milk

Suppose we observe the following purchase baskets over a period:

- B_1: {Coffee, Cream, Sugar}
- B_2: {Bread, Butter, Jam}
- B_3: {Cereal, Milk}
- B_4: {Coffee, Sugar}
- B_5: {Bread, Jam}

Step 2: Initialize Embedding Matrices

We use a simplified 2-dimensional embedding space for this example. In practice, the dimensionality would be much higher (e.g., 100-dimensional as in our full model). We initialize two 8×2 matrices randomly: U (output matrix) and V (input matrix). One initialization of U and V is shown in Table A1.

Table A1: Initial Embeddings

	$\mathbb{U}$ (output)		$\mathbb{V}$ (input)	
	Dim1	Dim2	Dim1	Dim2
Coffee	0.2	0.7	0.5	0.3
Cream	0.2	0.4	0.6	-0.2
Sugar	0.1	0.6	-0.1	0.8
Bread	0.3	0.7	0.3	0.5
Butter	0.8	-0.4	-0.7	-0.1
Jam	0.1	0.5	0.2	0.4
Cereal	-0.6	0.2	0.5	-0.6
Milk	0.4	-0.3	-0.3	0.3

Step 3: Training Process

For each basket, we update the embeddings as described in Section 3. We walk through one iteration for B_1: {Coffee, Cream, Sugar}. We will update the embeddings when Coffee is the input and we want to predict Cream and Sugar. These conditional purchase probabilities are given by:

$$P(\text{Cream}|\text{Coffee}) = \frac{\exp(u_{\text{cream}} \cdot v'_{\text{coffee}})}{\sum_k \exp(u_k \cdot v'_{\text{coffee}})}$$
$$P(\text{Sugar}|\text{Coffee}) = \frac{\exp(u_{\text{sugar}} \cdot v'_{\text{coffee}})}{\sum_k \exp(u_k \cdot v'_{\text{coffee}})}$$

Step 3a: Calculate Initial Probabilities

For ease of explanation, we use one negative sample: Bread. With Coffee as input (v'_{coffee}), we compute dot products with output vectors (u) of Cream, Sugar, and Bread as follows:

$$u_{\text{cream}} \cdot v'_{\text{coffee}} = (0.2 \times 0.5) + (0.4 \times 0.3) = 0.22$$
$$u_{\text{sugar}} \cdot v'_{\text{coffee}} = (0.1 \times 0.5) + (0.6 \times 0.3) = 0.23$$
$$u_{\text{bread}} \cdot v'_{\text{coffee}} = (0.3 \times 0.5) + (0.7 \times 0.3) = 0.36$$

Applying the `sigmoid` function to convert the dot products to probabilities, we get:

$$P(\text{Cream}|\text{Coffee}) = \frac{\exp(0.22)}{1 + \exp(0.22)} = 0.555$$
$$P(\text{Sugar}|\text{Coffee}) = \frac{\exp(0.23)}{1 + \exp(0.23)} = 0.557$$
$$P(\text{Bread}|\text{Coffee}) = \frac{\exp(0.36)}{1 + \exp(0.36)} = 0.589$$

Step 3b: Compute Log Loss

$$
\begin{aligned}
L &= \log P(\text{Cream}|\text{Coffee}) + \log P(\text{Sugar}|\text{Coffee}) + \log(1 - P(\text{Bread}|\text{Coffee})) \\
&= \log(0.555) + \log(0.557) + \log(1 - 0.589) \\
&= -0.588 - 0.585 - 0.889 = -2.063
\end{aligned}
$$

Step 3c: Update Embeddings

Using a learning rate $\eta = 0.1$, we update the embeddings to reduce the loss. The gradient update moves the embeddings of Coffee, Cream, and Sugar closer together while pushing Coffee and Bread apart. After this update step, the embeddings are shown in Table A2.

Table A2: Updated Embeddings After One Step

	$\mathbb{U}$ (output)		$\mathbb{V}$ (input)	
	Dim1	Dim2	Dim1	Dim2
Coffee	0.2	0.7	**0.495**	**0.303**
Cream	**0.222**	**0.413**	0.6	-0.2
Sugar	**0.122**	**0.6132**	-0.1	0.8
Bread	**0.270**	**0.682**	0.3	0.5
Butter	0.8	-0.4	-0.7	-0.1
Jam	0.1	0.5	0.2	0.4
Cereal	-0.6	0.2	0.5	-0.6
Milk	0.4	-0.3	-0.3	0.3

Step 4: Compute Product Similarities

To see the model's learning, we can compare the cosine similarities between Coffee and other products before and after the update. Note that in the full model, cosine similarity is computed using the output embeddings. However, in this illustrative example, we compute the cosine similarity between the updated *input* embedding for Coffee and the *output* embeddings for the other products. This is because here we only consider a single basket in the training data, in which only Coffee undergoes an input embedding update. Analogously, the output embeddings of the three other products (Cream, Sugar, and Bread) are updated. After the full training cycle, we can use the output embeddings to compute the similarity.

The cosine similarity between Coffee and Cream before the gradient update is:

$$
\begin{aligned}
c^{init}_{\text{coffee,cream}} &= \frac{v'_{\text{coffee}} \cdot u_{\text{cream}}}{||v'_{\text{coffee}}||||u_{\text{cream}}||} \\
&= \frac{(0.5 \times 0.2) + (0.3 \times 0.4)}{\sqrt{0.5^2 + 0.3^2}\sqrt{(0.3)^2 + 0.4^2}} \\
&= 0.843
\end{aligned}
$$

We can similarly compute the cosine similarities between Coffee and all other products after the update. The similarities are shown in Table A3.

Table A3: Cosine Similarities with Coffee

Product	Before Update	After Update
Cream	0.843	0.864
Sugar	0.648	0.678
Bread	0.811	0.800

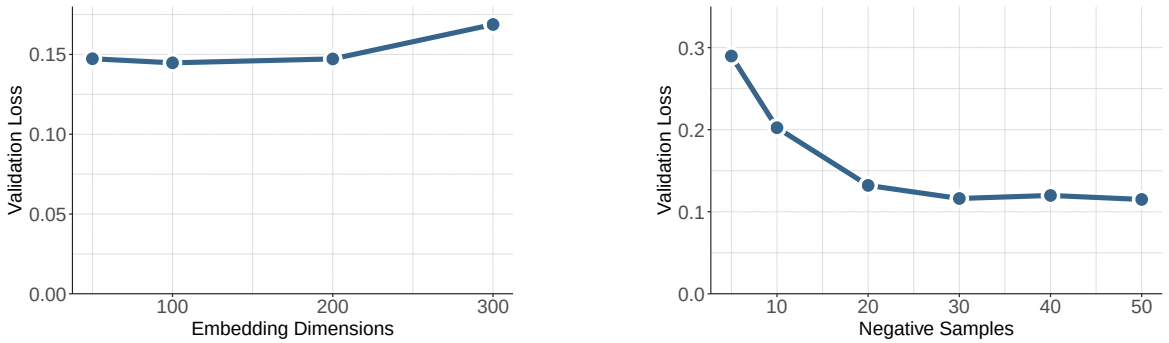
We can observe that after just one update step, the similarities between Coffee and its potential complements (Cream, Sugar) have increased. Correspondingly, the similarities with products not involved in the update remained unchanged.

This walkthrough presents the key steps of the embedding process: initializing matrices, iterating through purchase baskets, updating embeddings, and using the final embeddings to identify potential complements. It provides an intuitive understanding of how the model captures co-purchase patterns in a low-dimensional space, allowing for efficient discovery of product relationships even in large assortments.

A.2 Embeddings Model Training

We train the embeddings model described in Section 3 in Tensorflow using the Adam optimizer (Kingma and Ba, 2015). The training is done separately for purchase baskets and consideration sets. In both cases, we only consider product baskets with more than one product since our model is designed for within-basket/session relationships among products. Our entire sample is from January 2018 to June 2018. We train the model on data from January 2018 to May 2018, and leave June 2018 as test data. Within the training data, we keep a random 10% validation sample for hyperparameter tuning.

The key hyperparameters to tune are the embedding size $\mathcal{D}$ and the number of negative samples N_s. To get these parameters, we first do a random search by running a smaller version of the model using product categories (level 3 hierarchy) to identify a narrower range of hyperparameters over which the model performs well. This helps navigate the hyperparameter space efficiently and iterate quickly over different configurations of the model (Bergstra and Bengio, 2012). We then run the complete model using the actual products on this narrower range to learn the optimal hyperparameter values.



(a) Embedding Dimension

(b) Number of Negative Samples

Figure A1: Validation Loss with Model Hyperparameters

Typically, a larger dimension size $\mathcal{D}$ increases the complexity of the model as more embedding dimensions are able to capture product characteristics more accurately. However, there is a potential for over-fitting. We pick the optimal value of $\mathcal{D}$ based on loss on the validation sample. Analogously, more negative samples N_s help approximate the likelihood from Equation 4, and hence the gradient, more accurately. However, increasing the number of negative samples N_s substantially increases the cost of computing the likelihood, which severely affects the model's estimation time. For instance, increasing the number of negative samples from 10 to 20 doubles the training time. Hence, the trade-off here is balancing accuracy with model run time. Figure A1 shows how the validation loss varies with these parameters.

Further, in our framework, we set the dimensionality of both the purchase and consideration embeddings to be equal. Although in principle, the dimensionalities of the two spaces need not be identical, aligning them provides practical benefits. First, while the two embeddings capture different aspects of consumer behavior, enforcing an equal dimensionality simplifies the model architecture and reduces the number of hyperparameters, thereby streamlining model training and evaluation. Second, using a common embedding dimension facilitates subsequent operations—such as taking inner products or applying linear transformations—which are crucial for computing relationship scores (e.g., complementarity and substitutability) and, potentially, for combining the two embedding spaces in future extensions. This approach is consistent with common practice in the industry, where systems such as recommender engines and natural language processing pipelines typically employ uniform embedding dimensions for user and item representations to enable efficient similarity computations (Grbovic and Cheng, 2018).

A.3 Model Performance

Table A4 shows the model's performance on an out-of-time hold-out test set, which consists of all shopping sessions in June 2018. We test the model on purchase baskets and consideration sets separately, and in both cases, our product embeddings perform better. We compare the model's performance to popular benchmarks used in recommendation systems and market-basket analysis, including a baseline that uses only raw historical co-purchase/co-view rates.

Table A4: Model Performance with Hit Rate @ 10 on Held-Out Test Data

	Purchases	Consideration sets
Our product embeddings	**0.039**	**0.023**
LDA	0.035	0.018
SVD	0.033	0.021
Gensim	0.034	0.022
Non-negative matrix factorization	0.034	0.016
Shopper*	0.033	0.013
Exponential Family Embeddings*	0.031	0.022
Item-based collaborative filtering	0.029	0.020
Historical co-purchase/co-view rate	0.027	0.014

Note 1: The table shows performance for product embeddings and benchmark models on held-out test data as measured by hit-rate@10. The benchmark models were chosen based on a literature review and include the common approaches used in market-basket analysis and recommender systems. Hit rate has been rescaled by multiplying all values by 100. *Note 2:* * We were unable to train these models on the full data. To facilitate comparison, we subsampled 20% of the training data to train Shopper and Exponential Family Embeddings. The test data is the same for all models and was *not* subsampled.

A.4 Identifying Relationships for New Categories

In addition to testing relationships among products from different but pre-existing categories (typically created by the retailer), we can also generate new sub-categories of products and check how well they go with products from other categories. For instance, in Figure A2, we compare organic groceries with snack foods. Although there is no pre-defined organic category of products, as a proof-of-concept, we do a simple string search of the word "organic" in the names of the products. We then visualize them along with snack foods to see what kind of organic products are related to snack foods. The upper highlighted portion of Figure A2 shows a high degree of complementarity among nuts, dried seeds such as watermelon seeds, trail mixes, jerky and dried meats, and seaweed snacks. On the other hand, the lower highlighted portion of the graph shows less overlap and mainly consists of cookies, and chips and pretzels. We believe that having a flexible and scalable model such as this can provide crucial insights into market structure, brand competition, product positioning, user preferences, and personalized recommendations.

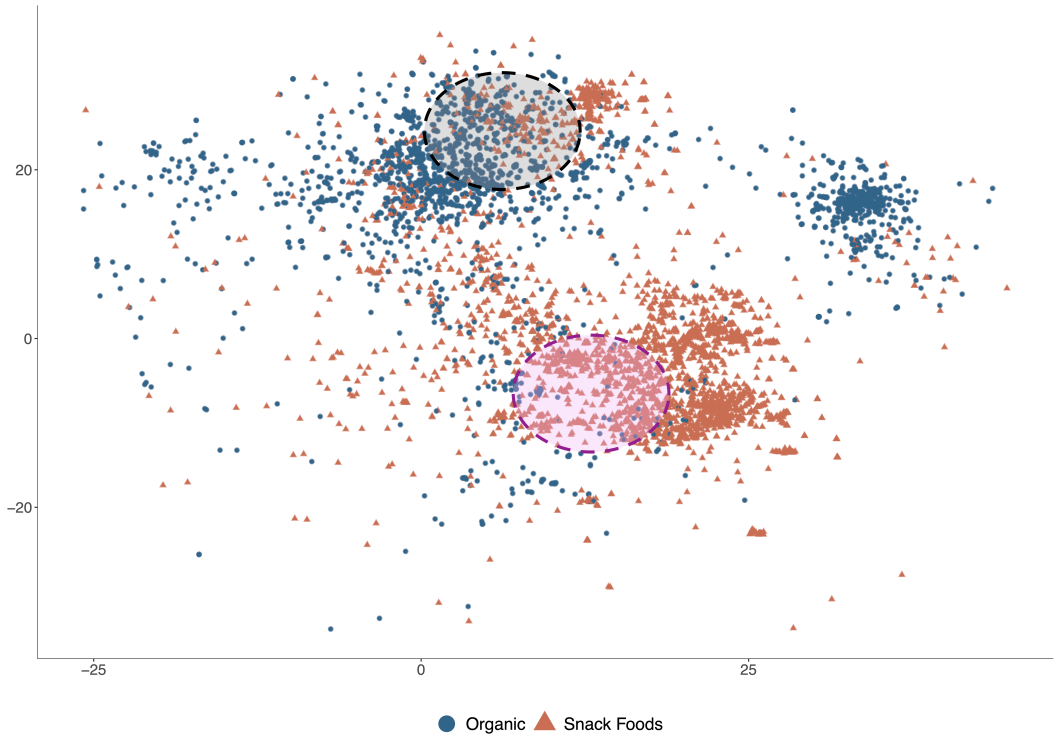


Figure A2: Purchase embeddings for organic groceries and snack foods

B Product Similarity Based on Embeddings

We provide examples of products based on their proximity to a focal product in the purchase and consideration set spaces. In Table B5, we consider *Organic Russet Potatoes*. The top panel shows the top 10 products closest to it in the purchase space, along with their embeddings-based complementarity scores. We find that the products closest to organic potatoes in the purchase space include other organic fruits and vegetables such as organic celery, organic grape tomatoes, and organic green bell peppers. Similar to the purchase space, we inspect the product closest to *Organic Russet Potatoes* in the consideration space in the bottom panel of Table B5. We see that organic potatoes are now closer to other varieties of potatoes in the consideration space, such as golden potatoes, red potatoes, and even sweet potatoes, indicating a higher degree of substitutability among them. This is in contrast to the purchase space, where organic potatoes were closer to other organic fruits and vegetables.

Table B5: Products Closest to *Organic Russet Potatoes* in the Purchase and Consideration Space

Organic Russet Potatoes, 5 Lb (10-12 Ct)	
Purchase Space	
Product	Comp. Score
Organic Celery Hearts, 16 Oz	0.75
Organic Grape Tomatoes, 1 Pint	0.74
Organic Green Bell Peppers, 2 Ct	0.74
Organic Carrots, 2 Lb	0.73
Organic Cauliflower, 1 Ct	0.73
Organic Garlic, 8 Oz	0.72
The Farmers Hen Large Organic Eggs, 1 Dozen	0.70
Organic Bananas, Minimum 5 Ct	0.69
Organic Broccoli Crowns, 2 Ct	0.69
Organic Romaine Hearts, 3 Ct	0.69
Consideration Set Space	
Product	Sub. Score
Green Giant Organic Golden Potatoes, 3 Lb	0.95
Green Giant Organic Red Potatoes, 3 Lb	0.95
Organic Russet Potatoes, 3 Lb	0.95
Green Giant Klondike Gourmet Petite Purple-Purple Fleshed Potatoes, 24 Oz	0.93
Green Giant Klondike Fingerling Potatoes, 24 Oz	0.93
Green Giant Golden Potatoes, 5 Lb	0.92
Organic Sweet Potatoes, 3 Lb	0.92
Green Giant Klondike Petite Red-White Fleshed Potatoes, 24 Oz	0.92
The Little Potato Garlic Herb Potato Microwave Kit, 16 Oz	0.92
The Little Potato Company Garlic Herb Oven Griller Kit, 16 Oz	0.91

Note 1: The complementarity score (Comp. score) is a measure of proximity in the purchase space and is indicative of complementarity. A higher score implies stronger potential complementarity. The substitution score (Sub. score) is a measure of proximity in the consideration space and is indicative of substitutability. A higher score implies stronger potential substitutability.

In Tables B6 and B7, we provide examples for two more focal products from different categories – 1) dish-washing liquid from cleaning supplies, and 2) oil-free acne face wash from skin care, respectively. For example, for the dish-washing liquid, other household and cleaning items such as paper towels, laundry detergents, and steel cleaners are closer in the purchase space. In the consideration space, dish-washing liquid is now closer to other types/brands of dish-washing detergents, such as liquids and soaps of different scents and sizes.

In the case of oil-free acne face wash shown in Table B7, products that go along with this facial cleanser include other hygiene and beauty products such as liners, rash cream, body wash, and deodorant. In the consideration space, the acne face wash shows considerable similarity with varieties of acne face washes.

Table B6: Products Closest to *Dishwashing Liquid* in the Purchase and Consideration Space

Joy Ultra Dishwashing Liquid, Lemon Scent, 12.6 oz	
Purchase Space	
Product	Comp. Score
Bounty Paper Towels, White, 12 Super Rolls	0.50
Tide PODS Plus Downy HE Turbo Laundry Detergent Pacs	0.48
P&G 45Oz Cmp Gel Detergent	0.48
The Art of Shaving Shave Cream, Sandalwood	0.47
Bounty Towel, Bounty Essentials	0.46
Bounty Paper Towels, Select-A-Size, 6 Triple Rolls	0.46
Lillian Dinnerware Pebbled Plastic Plate	0.45
Saratoga Spring Water	0.45
CLR Stainless Steel Cleaner	0.45
Pepcid Complete Dual Action Acid Reducer and Antacid Chewcap	0.45
Consideration Set Space	
Product	Sub. Score
Joy Dishwashing Liquid, Lemon, 5gal Pail	0.83
Joy Dishwashing Liquid 38 oz Bottle	0.79
Joy Dishwashing Liquid Lemon Scent 12.6 oz Bottle	0.71
Palmolive Ultra Anti-Bacterial Dish Soap, Orange, 56 Oz	0.70
Ajax Triple Action Dish Soap, Orange, 12.6 Oz	0.69
Palmolive Ultra Dish Soap, Orange, 25 Fl Oz	0.69
Biokleen Natural Dish Liquid, Citrus, 32 Oz, 12 Ct	0.69
Palmolive OXY Plus Power Degreaser Dish Soap, 10 Oz	0.69
Ajax Super Degreaser Dish Soap, Lemon, 52 Oz	0.69
Ajax Dish Soap, Tropical Lime Twist, 52 Oz	0.68

Note 1: The complementarity score (Comp. score) is a measure of proximity in the purchase space and is indicative of complementarity. A higher score implies stronger potential complementarity. The substitution score (Sub. score) is a measure of proximity in the consideration space and is indicative of substitutability. A higher score implies stronger potential substitutability.

Table B7: Products Closest to *Oil-Free Acne Wash* in the Purchase and Consideration Space

Neutrogena Oil-Free Acne Wash Redness Soothing Cream Facial Cleanser, 6 Fl. Oz	
Purchase Space	
Product	Comp. Score
U By Kotex Barely There Daily Liners	0.53
Aveeno Active Naturals Daily Moisturizing Body Yogurt Body Wash	0.52
Palmer's Cocoa Butter Formula Bottom Butter	0.52
Neutrogena Oil-Free Acne Wash Redness Soothing Facial Cleanser	0.51
Neutrogena Oil-Free Acne Face Wash Pink Grapefruit Foaming Scrub	0.49
Maybelline New York Fit Me Matte & Poreless Foundation, Natural Beige	0.48
Secret Invisible Solid Anti-Perspirant Deodorant 2 Ct	0.47
Motrin IB, Ibuprofen, Aches and Pain Relief	0.47
Nature's Bounty Hair, Skin & Nails Gummies Strawberry	0.47
Equate Ibuprofen Pain Reliever/Fever Reducer 200 mg Tablets	0.46
Consideration Set Space	
Product	Sub. Score
Neutrogena Oil-Free Acne Face Wash With Salicylic Acid, 6 Oz.	0.84
Neutrogena Oil-Free Acne Face Wash Daily Scrub With Salicylic Acid, 4.2 Fl. Oz.	0.84
Neutrogena Oil-Free Acne Face Wash Pink Grapefruit Foaming Scrub	0.83
Neutrogena Naturals Purifying Pore Scrub, 4 Fl. Oz.	0.82
Neutrogena Rapid Clear Stubborn Acne Cleanser, 5 Oz	0.82
Neutrogena All-In-1 Acne Control Daily Scrub, Acne Treatment 4.2 Fl. Oz.	0.82
Neutrogena Oil-Free Acne Wash Pink Grapefruit Cream Cleanser, 6 Oz	0.82
Neutrogena Oil-Free Acne Face Wash With Salicylic Acid, 9.1 Oz.	0.81
Neutrogena Men Oil-Free Invigorating Foaming Face Wash, 5.1 Fl. Oz	0.80
Neutrogena Oil-Free Acne Face Wash Pink Grapefruit Foaming Scrub, Salicylic Acid Acne Treatment, 6.7 Fl. Oz.	0.80

Note 1: The complementarity score (Comp. score) is a measure of proximity in the purchase space and is indicative of complementarity. A higher score implies stronger potential complementarity. The substitution score (Sub. score) is a measure of proximity in the consideration space and is indicative of substitutability. A higher score implies stronger potential substitutability.

C Experiment Design Details

We provide more details on the experiment design here. Basic characteristics of the bundles created by the four strategies are shown in Table C8. We show the results for $11,296$ bundles which were viewed at least once during the experiment and hence are part of our subsequent analysis. All values in this table are calculated based on the pre-experiment data used for training. The number of bundles viewed is different across the four bundle types since there is flux in the inventory, and depending upon the location and time of the consumer, a bundle may or may not be available. We calculate the co-purchase rate for the product pairs. *Price-1*, *Rating-1*, *Purchase rate-1* correspond to the average price of the focal products, their average user-provided rating, and their historical individual purchase rates. Analogously, *Price-2*, *Rating-2*, and *Purchase rate-2* correspond to the same variables for the add-on product. The last four rows show the mean of binary variables, which take the value 1 if both products belong to the same department, aisle, category, and brand, respectively.

Table C8: Mean pre-experiment product characteristics for candidate bundles in the experiment

	Cross category	Cross dept.	Variety	Co-purchase
	(CC)	(DC)	(VR)	(CP)
Bundles	3,092	2,418	3,287	2,499
Comp. score (c_{ij})	0.34	0.22	0.41	0.40
Sub. score (s_{ij})	0.39	0.20	0.74	0.58
Co-purchase rate	0.10	0.02	0.23	0.29
Price - 1	10.70	10.96	10.22	9.53
Price - 2	8.32	10.16	10.06	8.63
Purchase rate - 1	0.04	0.04	0.04	0.04
Purchase rate - 2	0.05	0.06	0.04	0.04
Product rating - 1	4.70	4.72	4.69	4.69
Product rating - 2	4.70	4.80	4.69	4.72
Same dept	0.99	0.00	1.00	0.88
Same aisle	0.55	0.00	0.99	0.71
Same category	0.02	0.00	0.98	0.53
Same brand	0.17	0.10	0.46	0.44

Note 1: Co-purchase rate has been multiplied by 100. Price-1, Purchase rate-1, and Product rating-1 show the average price, average historical purchase rate, and the average product rating of the focal product in each bundle type. Price-2, Purchase rate-2, and Product rating-2 are corresponding variables for the add-on product. Same department, Same aisle, Same category, and Same brand are binary variables that indicate if the two products are from the same department, aisle, category, and brand, respectively.

C.1 Balance Checks

Our experiment was at a user-product level, i.e., a user for a given product was only exposed to one of the four candidate bundles for that product throughout the experiment period. To verify that the

randomization happened correctly, we check for imbalance across the four experimental groups. The results from the balance checks are shown in Table C9. The last column shows the p-value from an ANOVA testing for equality of means across the arms. We do not find any evidence for imbalance.

Table C9: Pre-experiment summary statistics for key variables

Variable means and standard deviations

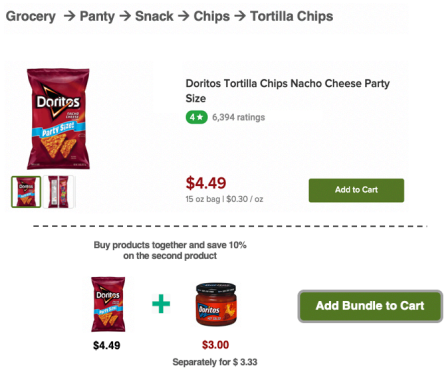
	Co-purchase	Cross-cat.	Cross-dept.	Variety	p-value
	(CP)	(CC)	(DC)	(VR)	
Observations	5,642	5,395	4,846	5,369	
Visits	3.28	3.28	3.15	3.35	0.25
	5.58	4.75	4.74	5.5	-
Product views	7.99	8.21	7.93	8.19	0.34
	9.61	9.88	9.59	9.78	-
Products ATC	0.35	0.35	0.35	0.32	0.71
	0.75	0.7	0.75	0.64	-
Units purchased	0.49	0.47	0.48	0.47	0.25
	1.77	1.58	1.72	1.74	-
Revenue	4.71	4.56	4.72	4.51	0.28
	19.26	19.32	20.03	18.09	-

Note 1: Pre-experiment means and standard deviations for users who visited the retailer's website at least once during the month prior to the start of the experiment.

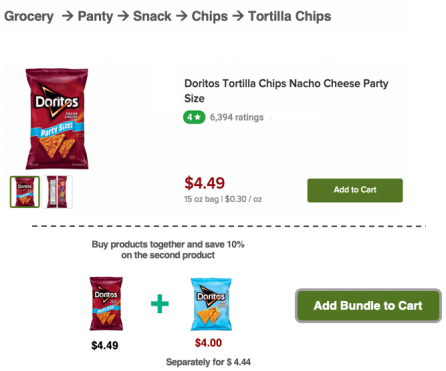
Note 2: Product views include all page visits by the user. Products ATC accounts for the total number of different products added-to-cart whereas units purchased account for multiple units of the same product bought.

Note 3: p-values are from a test of equality of means (ANOVA).

C.2 Example Bundle Recommendations Shown in the Experiment



(a) Complementary bundle recommendation example



(b) Variety bundle recommendation example

Figure C3: Illustrative Examples from the Field Experiment

C.3 Relative Performance in Each Experiment Arm

In the field experiment, each focal product has four bundle recommendations associated with it, i.e., it has four treatment arms – a) Cross-category complement, b) Cross-department complement, c) Variety, and d) Co-purchase. Recall that our goal from the experiment is *not* to identify which bundle recommendation strategy has the best performance on average, but rather to learn which treatment/recommendation pair is best for each product in order to maximize revenue from these recommendations. Furthermore, the design of the experiment forced some of the arms to select distinct products, such that these arms do not uniformly correspond to the top-scoring bundle recommendation by that arm's selection rule; thus, these comparisons are not directly informative even about the average performance of these bundle recommendation strategies. While this is not central to our analysis, we report here a naive comparison of the relative performance of the four experiment arms for the sake of completeness.

We consider the co-purchase recommendation as the baseline and compare the performance of the remaining treatments relative to it. Tables C10 and C11 report the relative performance of the experiment arms on add-to-cart and revenue. Column 1 in both tables compares the co-purchase strategy with the cross-category complement strategy, Column 2 compares co-purchase with cross-department complements, and Column 3 compares it with the variety strategy. In all columns, the standard errors are clustered at the user and focal product level.

Table C10: Naive Comparison of Add-to-Cart Rates in Experimental Arms versus the Co-Purchase Arm

Dependent Variable:		ATC	
Model:	(1)	(2)	(3)
Variables			
(Intercept)	0.0140***	0.0140***	0.0140***
	(0.0009)	(0.0009)	(0.0009)
Cross Cat. Complement	-0.0060***		
	(0.0009)		
Cross Dept. Complement		-0.0095***	
		(0.0009)	
Variety			-0.0019*
			(0.0011)
Fit statistics			
Observations	118,500	110,591	118,170
R^2	0.00054	0.00142	3.89×10^{-5}

Clustered (User & Focal Product) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table C11: Naive Comparison of Revenue in Experimental Arms versus the Co-Purchase Arm

Dependent Variable:	Revenue		
Model:	(1)	(2)	(3)
Variables			
(Intercept)	0.1200***	0.1200***	0.1200***
	(0.0126)	(0.0126)	(0.0126)
Cross Cat. Complement	-0.0723***		
	(0.0132)		
Cross Dept. Complement		-0.0990***	
		(0.0129)	
Variety			-0.0155
			(0.0177)
Fit statistics			
Observations	118,500	110,591	118,170
R^2	0.00027	0.00054	6.81×10^{-6}

Clustered (User & Focal Product) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

D Predictive model for learning the optimized policy

We compare several predictive models to select the best for learning the optimized policy. Table D12 shows the out-of-sample performance of different classes of classification models used to predict bundle recommendation add-to-cart (ATC), i.e., if a user adds the recommended pair to their cart. We use an up-the-funnel surrogate, ATC, as the dependent variable since it provides more power as compared to fitting the model directly on purchases, which is a highly imbalanced target variable. Recent work on surrogate metric learning has indicated that training policies on surrogates could be a valuable alternative even if we observe the desired downstream outcome, since the surrogates tend to have less variance and more power (Yang et al., 2024; Gordon et al., 2023). In our comparative exercise, we find that XGBoost Chen and Guestrin (2016) outperforms other models and hence we select it as our outcome model for policy learning.

Table D12: AUC for predicting add-to-cart on hold-out test data

Model	AUC
Baseline	0.6635
Logistic	0.7023
Hierarchical Logistic	0.6967
LASSO	0.7009
Random Forest	0.7073
XGBoost	**0.7276**

Note 1: The baseline model excludes the product relationship heuristics – c_{ij} and s_{ij}

Note 2: Hierarchical Logistic regression is a mixed effects model with varying slopes that allows the effects of c_{ij} and s_{ij} to vary by product category. We estimate it using Restricted Maximum Likelihood.

XGBoost Outcome Model

The optimal hyperparameters of the selected XGBoost model are shown in Table D13a, and the pre-treatment covariates used are shown in Table D13b. We also show sample trees from the outcome model in Figure D4.

Table D13: XGBoost Model: Hyperparameters and Features

(a) Optimal Hyperparameters for XGBoost

Parameter	Value
Learning rate (η)	0.06
Number of trees (`nrounds`)	100
Maximum depth (`max_depth`)	3
Subsample observations (`subsample`)	0.6
Subsample features (`colsample_bytree`)	0.5
Minimum instances for child node (`min_child_weight`)	2
L1 regularization (`reg_alpha`)	0.003

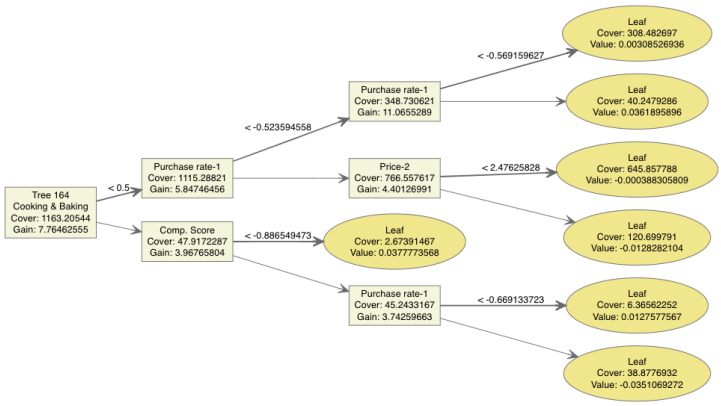
Note 1: This table shows the optimal hyperparameters for the XGBoost model used to learn the optimized bundling policy. The hyperparameters were chosen using a 10-fold cross-validation.

(b) Pre-treatment covariates

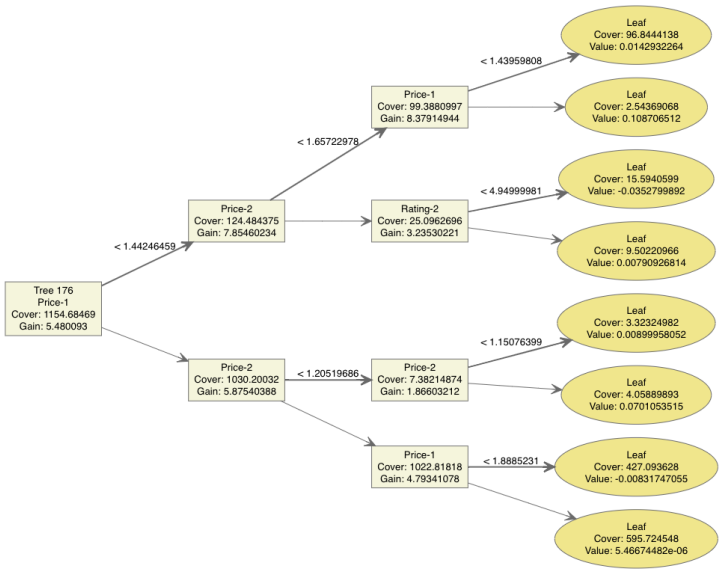
No.	Feature
1.	Complementarity score
2.	Substitutability score
3.	Historical co-purchase rate
4.	Price of the focal product
5.	Price of the add-on product
6.	5-scale rating of the focal product
7.	5-scale rating of the add-on product
8.	Historical purchase rate of the focal product
9.	Historical purchase rate of the add-on product
10.	Indicator for same brand
11.	Indicator for same aisle but different category

Note 1: This table shows pre-treatment features used in the outcome model.

Sample trees from XGBoost model used to optimize the bundling policy



(a) Sample tree showing the interaction between the complementarity score and focal product aisle



(b) Sample tree showing the interaction between prices of the two products

Figure D4: Sample trees from the XGBoost model trained to learn the optimized bundle design policy

E Robustness Checks for Optimized Policy

We perform several robustness checks for our learned policy, including: 1) extending it to all focal products in the experiment, 2) comparing the value of the policy relative to two baselines, historical co-purchases and co-views, and 3) computing the value-added by the embeddings over and above the granular clickstream data.

Optimized Policy Using All Focal Products

In Figure 5, we use 897 focal products for which all four bundle recommendation strategies had at least one view during the experiment. Here, we repeat the analysis using all the focal products. We learn the optimized bundling policy in the same way as before. For focal products that did not have all four bundle recommendations viewed, the optimized policy selects from the remaining set. The results are shown in Figure E5. The policy that uses both scores performs the best here as well. Moreover, the results for this policy are qualitatively similar to the policy optimized in Figure 5.

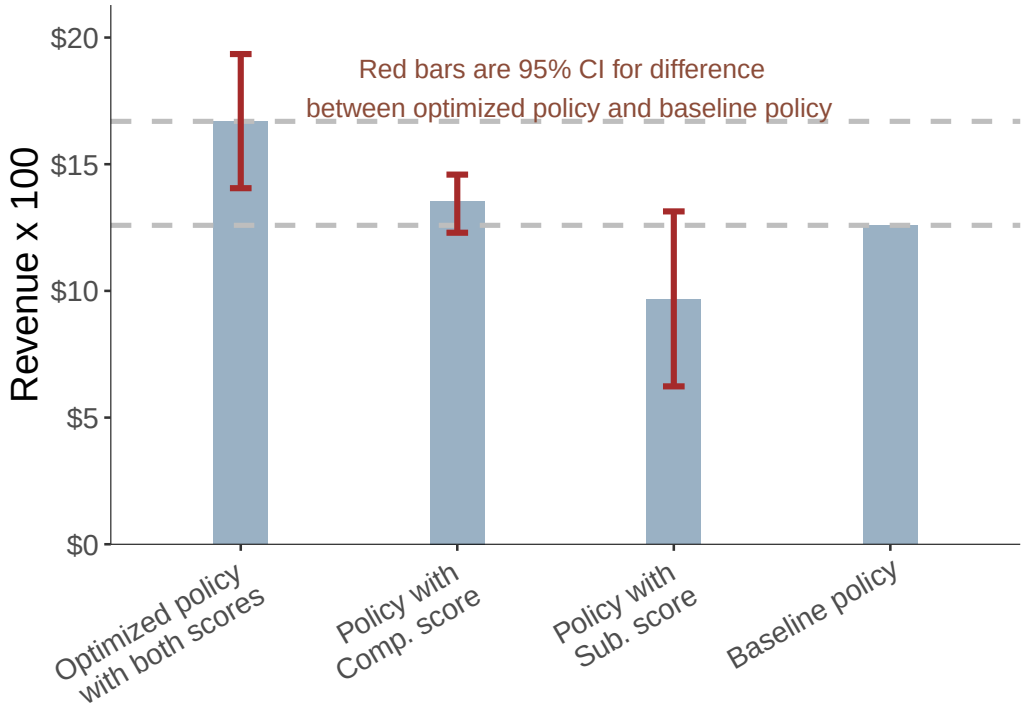


Figure E5: Revenue from the Bundle Recommendation Policy Optimized using All Focal Products

Note: The optimized policy uses both scores to find the best bundle recommendation for a focal product. The second bar only uses the complementarity score to find the best bundle recommendations, and the third bar uses the substitutability score and selects the variety bundle recommendation. The last bar is the benchmark policy-based historical co-purchases.

Data vs. Embeddings Model

We extend the scope of our main analysis shown in Figure 5 by introducing another baseline policy – historical co-views – that selects the bundle recommendation based on the historical co-views of products. Recall that while this co-view strategy was not part of our original field experiment, our off-policy evaluation framework enables us to compare *any* counterfactual policy. However, since co-views are sparse and the policy was not directly used in the experiment, we need to modify the bundle recommendation selection criterion to include this policy in the analysis. Specifically, we adopt a thresholding policy where a bundle recommendation pair with 25 co-views is selected as part of the co-view policy. This gives us a reasonable set of purchases over which we can evaluate and compare the co-view policy with other policies.

The results are shown in Figure E6, and we see that the optimized bundle recommendation policy performs well against all other policies. Further, juxtaposing the co-purchase and co-view policies with the optimized policy emphasizes the relative contribution from our embedding-based approach vs. only using granular clickstream data.

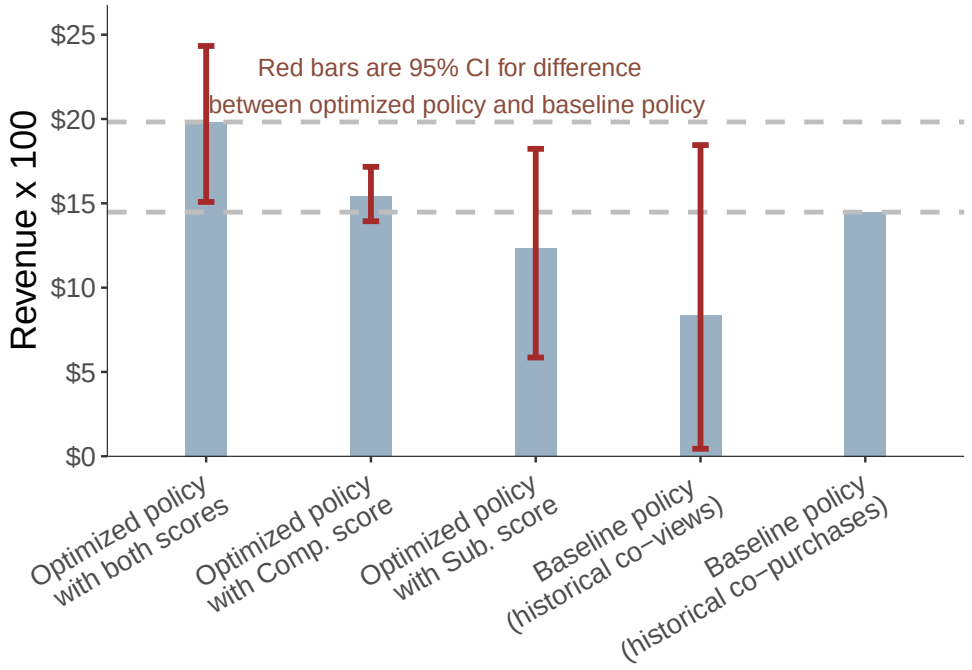
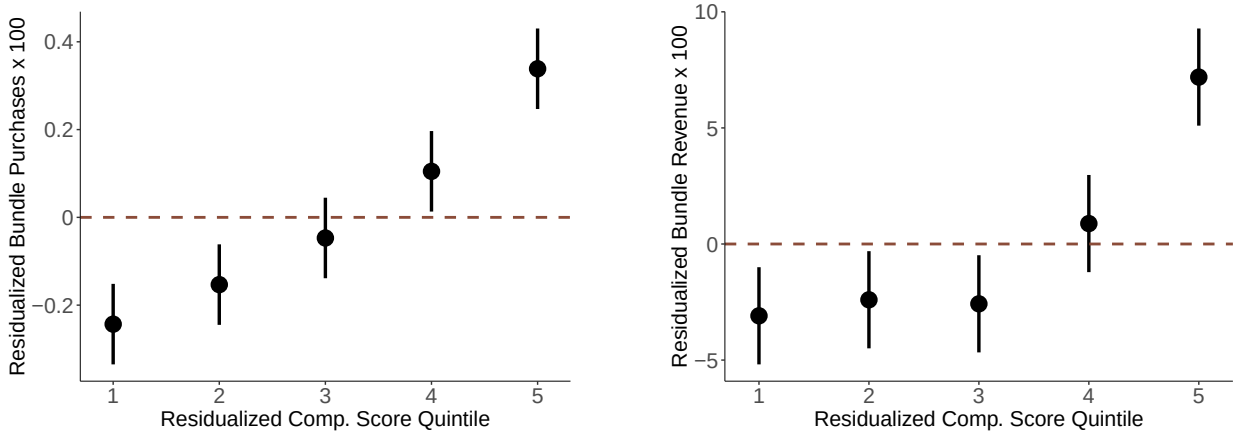


Figure E6: Revenue from the Optimized Bundle Recommendation Policy vs. Policies that Only Use Either the Potential Complementarity Score (c_{ij}) or the Potential Substitutability Score (s_{ij})

Note: The optimized policy is trained on 897 focal products for which all bundle recommendations received at least one view during the experiment. It uses both scores to find the best bundle recommendation for a given focal product. The second bar only uses the complementarity score to find the best bundle recommendation, and the third bar only uses the substitutability score, i.e., the variety bundle recommendation. The fourth bar uses historical co-views, and the last bar is the benchmark policy-based historical co-purchases.

Effect of Product Relationship Scores Net of Co-purchases and Co-Views

Figure E7 shows the results from the exercise described in Section A.2, where we test the value of recommendation pairs used in the field experiment from a different perspective. We assess the value-added by the product relationship scores, specifically the complementarity score (c_{ij}) over and above the historical co-purchase rate. We regress residualized bundle success (purchases and revenue) on residualized complementarity score (c_{ij}). We residualize both variables by regressing them on the historical co-purchase rate. The left panel of the figure shows the effect on residualized bundle purchases by residualized complementarity score quintiles. The right panel shows the impact on bundle revenue. In both cases, we see a large increase in bundle success as the complementarity score increases.

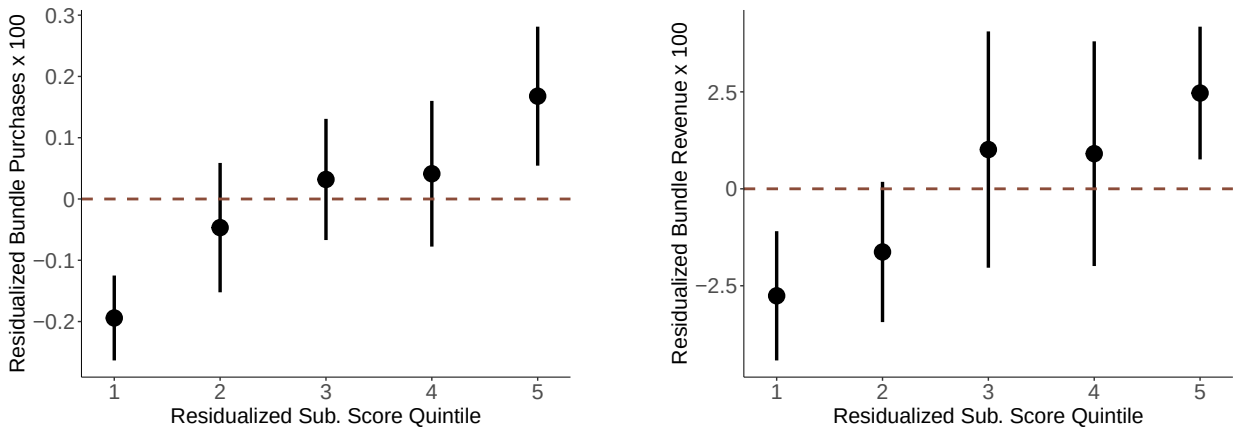


(a) Effect on Purchases

(b) Effect on Revenue

Figure E7: Effect of Potential Complementarity Score on Bundle Recommendation Success Net of Historical Co-purchases

We repeat the exercise for the potential substitutability score by netting out the effect of historical co-views. The results are qualitatively similar to the ones above and are shown in Figure E8.



(a) Effect on Purchases

(b) Effect on Revenue

Figure E8: Effect of Potential Substitutability Score on Bundle Recommendation Success Net of Historical Co-views

F Hierarchical model

To build intuition on how the product relationship scores influence bundle success and also to check their robustness across categories, we build a hierarchical logistic regression with varying slopes. We allow the effects of c_{ij} and s_{ij} to vary by the focal product aisle and the intercepts to vary by both the focal product aisle and the add-on product aisle. We estimate the model shown in Equation 1.

$$Pr(Y_{ij} = 1|\text{View}_i) = \text{logit}^{-1}\left(\alpha_{k[i]j} + \alpha_{k[j]i} + \beta^c_{k[i]j} c_{ij} + \beta^s_{k[i]j} s_{ij} + \gamma^T W_{ij}\right) \tag{1}$$

$$\alpha_{k[j]} \sim \mathcal{N}(0, \sigma^2)$$

$$\begin{pmatrix} \alpha_k[i] \\ \beta^c_k[i] \\ \beta^s_k[i] \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_\alpha \\ \mu_{\beta^c} \\ \mu_{\beta^s} \end{pmatrix}, \Sigma\right)$$

where i indexes the focal product, j is the add-on product, $k[i]$ is the aisle of the focal product and $k[j]$ is the aisle of the add-on product. Together, ij make one bundle, and we estimate the probability of the bundle being added-to-cart, conditional on the user viewing the focal product i. The intercept α_k is allowed to vary by both aisles. The slopes β^c and β^s, i.e., the coefficients for the complementarity heuristic c_{ij} and the substitutability heuristic s_{ij} vary by the aisle of the focal product. Pre-treatment variables and other product meta-data are captured by the vector W with their parameters γ held fixed. The varying parameters are estimated jointly with each parameter having a separate mean and variance. Their variances and covariances are given by the 3×3 matrix Σ.

Setting up the model in a hierarchical fashion helps us account for unobserved variation in product aisles that is not captured in a pooled regression model. Further, since we model the product aisles separately, we can use the model to generate predictions for aisles that were not part of our training data, and hence generalize our findings to a larger domain. In addition to statistical benefit, the hierarchical modeling also provides a systematic way to analyze the robustness of our model across aisles, as we show in Figure 7.

Estimation results from Model 1 are shown in the first column of Table F14. We find that both scores are significant and positive, even after controlling for basic product features, including the historical co-purchase rate. In the second column, we test whether asymmetry in prices has an effect on bundle success. We include a binary indicator that takes the value 1 if Price-1 $\geq$ Price-2. We find that bundles where the add-on product has the same or a lower price are more successful. The last coefficient in column 2 shows the result. We also test this using ANOVA. The results are shown in Table F15.

Table F14: Mixed effects hierarchical logistic regression to predict bundle add-to-cart

	Dependent variable:	
	Bundle add-to-cart	
	(1)	(2)
Comp. score	0.377***	0.373***
	(0.046)	(0.046)
Sub. score	0.276***	0.265***
	(0.048)	(0.048)
Hist. co-purchase rate	0.068***	0.066***
	(0.010)	(0.010)
Price-1	−0.064	−0.154***
	(0.047)	(0.056)
Price-2	−0.350***	−0.251***
	(0.051)	(0.060)
Same brand	0.342***	0.330***
	(0.062)	(0.062)
Diff. category — Same aisle	0.212***	0.209***
	(0.059)	(0.059)
Hist. purchase rate-1	−0.014	−0.011
	(0.027)	(0.027)
Hist. purchase rate-2	0.027	0.026
	(0.027)	(0.027)
Rating-1	0.078	0.078
	(0.070)	(0.070)
Rating-2	−0.068	−0.067
	(0.063)	(0.064)
Price-1 $\geq$ Price-2		0.245***
		(0.077)
Observations	227,311	227,311
Log Likelihood	−10,706	−10,701
Bayesian Inf. Crit.	21,647	21,649

Note 1: *p<0.1; **p<0.05; ***p<0.01
Note 2: Intercepts vary by focal and add-on product aisles. c_{ij} and s_{ij} vary by focal product aisle.

Table F15: Wald's test for equality of two price coefficients in Model 1

	χ^2	Chi Df	Pr($>\chi^2$)
Price-1 = Price-2 in Model 1	10.44	1	0.001

G Consumer Survey: Design and Results

We posit that our embeddings-based heuristics are indicative of potential complementarity and potential substitutability between two products. Ideally, we would like to formalize this argument by linking these heuristics to an underlying cross-price elasticity matrix. However, it is infeasible to estimate such a matrix given the sparsity of co-purchases. In fact, it is this infeasibility that motivated us to seek alternative, scalable approaches in the first place.

Nevertheless, we provide external validation for our product relationship scores using online choice experiments that we run with participants on Prolific. We randomly sample 400 products from the retailer's assortment, which we call the focal products. For each focal product, we first sample a potential complement with probability proportional to the cosine similarity in purchase space. Then we sample another product with probability proportional to $1 -$ cosine similarity in the purchase space. We repeat the process for potential substitutes by sampling from the consideration space while keeping the 400 focal products fixed.

We then run a conjoint-style experiment on Prolific with 200 users where we show them a focal product and ask which of two options they are more likely to consume *along with* the focal product in case of potential complements. Analogously, for potential substitutes, we ask them which product they are more likely *to switch to* if the focal product is *out of stock.* In total, each user responds to 10 randomly selected focal products – 5 for potential complements and 5 for potential substitutes. The order in which the focal products are shown is random, and the order of the two options within each focal product is also random.

We compare the alignment between the responses from the survey and the predictions made by our product relationship heuristics, i.e., we test whether the consumers and the product relationship heuristics agree on which products are more likely to be potential complements and which products are more likely to be potential substitutes.

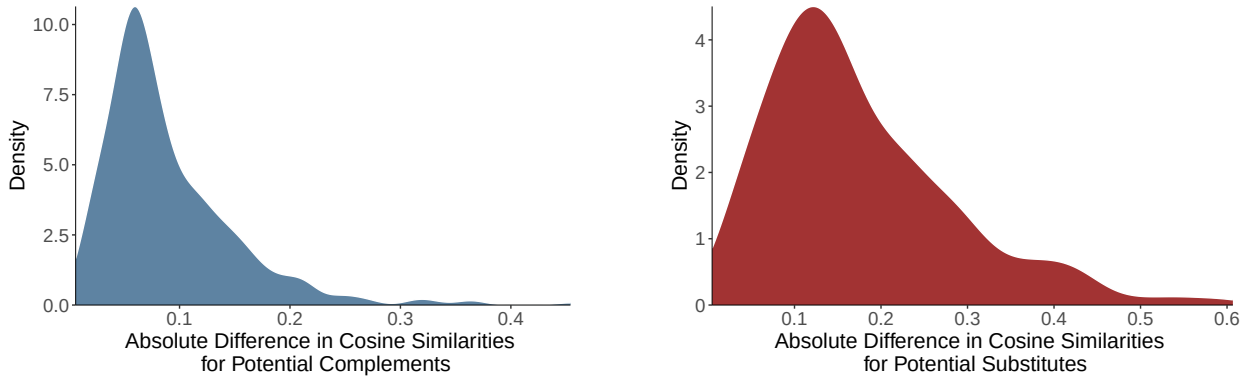
Table G16: Overview of Data from Consumer Survey

Number of Users	200
% Female	56.5
Number of Unique Products Shown	397
Mean Cosine Similarity - Potential Complements	0.091
Mean Cosine Similarity - Potential Substitutes	0.177

We begin the analysis of the consumer survey with Table G16, which presents an overview of the survey data. 200 users took the survey, and each user answered 10 questions – 5 for potential complements and 5 for potential substitutes. Out of the 400 products in the survey, 397 had at least one user response.

Recall that the survey is designed in a conjoint style, i.e., each user is shown two options for a particular product and they have to pick one. The key independent variable of interest is the difference in product relationship scores between the two options. The mean difference in these scores for po-

tential complements and potential substitutes is shown in the last two rows of Table G16, and the full distribution is shown in Figure G9.



(a) Potential Complements

(b) Potential Complements

Figure G9: Difference in Product Relationship Scores of Two Options

To check the alignment between the consumer responses and the product relationship scores, we regress the choice made by the user on the difference between the scores. A strong, positive, significant coefficient would indicate that high alignment between the scores and the consumer responses. This is exactly what we find. Table G17 shows the results for potential complements and potential substitutes separately. Column 1 regresses the consumer choice on the difference between the scores of the two products for potential complements, and Column 2 does it for potential substitutes. In both cases, we find that the difference in the scores is highly predictive of consumer choice. This means that the product pairs that have high complementarity scores are more likely to be picked for joint consumption by consumers. Similarly, product pairs that have high substitutability scores from the embeddings are more likely to be switched between by the consumers. Columns 3-6 replicate these results with respondent and category-fixed effects. In all models, we cluster the standard errors at the user level.

This evidence lends support to our claim that the embeddings-based heuristics from the purchase space are indicative of potential complementarity and those from the consideration space are indicative of potential substitutability.

Table G17: Regression Results from the Consumer Survey

Dependent Variable:	Choice (Binary)					
	Potential Complements	Potential Substitutes	Potential Complements	Potential Substitutes	Potential Complements	Potential Substitutes
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
(Intercept)	0.4773***	0.5271***				
	(0.0164)	(0.0131)				
Cosine Similarity Difference	0.9528***	1.198***	1.022***	1.215***	1.008***	1.213***
	(0.1235)	(0.0536)	(0.1408)	(0.0626)	(0.1409)	(0.0644)
Fixed-effects						
Respondent			Yes	Yes	Yes	Yes
Focal Product Category					Yes	Yes
Fit statistics						
Observations	1,000	1,000	1,000	1,000	1,000	1,000
R^2	0.04379	0.25352	0.25632	0.39167	0.26394	0.39848
Within R^2			0.05266	0.25705	0.05145	0.25498

Standard-errors clustered by respondent in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

H Extensions: Multi-product Bundle Recommendations

H.1 Multiple Bundle Recommendations for Each Focal Product

Our main analysis focuses on identifying the single best bundle recommendation for each focal product. However, as noted in Section 3.2, we create multiple candidate recommendations for each focal product during our field experiment. Here, we leverage this capability to extend our approach by generating and evaluating multiple bundle recommendations simultaneously for each focal product, rather than selecting just one optimal pairing. This extension aligns with common industry practices where retailers frequently present customers with several bundled product options. By offering multiple recommendations, retailers can mitigate the risk associated with recommending a single product pair and potentially increase the likelihood of matching diverse customer preferences.

To implement this extension, we utilize the same XGBoost model described in Section 7. For each focal product, we generate all possible bundle recommendation combinations with other products in the assortment and score them using our trained XGBoost model. We then rank these combinations by their predicted success probability and select the top four recommendations. This process yields a set of recommendations that vary in their composition, prices, and revenue potential, providing customers with diverse options that may better accommodate heterogeneous preferences.

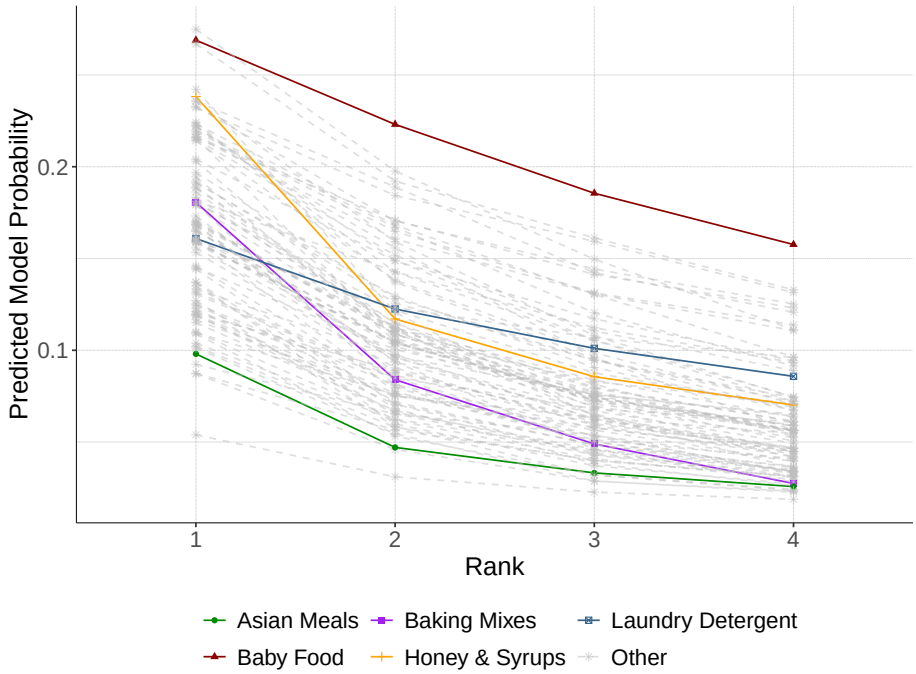


Figure H10: Heterogeneity in Probability of Success of Multiple Bundle Recommendations Across Product Categories

Figure H10 illustrates the heterogeneity in predicted probabilities across product categories. The figure displays the average predicted probability for the top four recommendations by rank, broken down by focal product category. We see that the rate of decline varies substantially across categories, indicat-

ing category-specific heterogeneity in recommendation quality. For instance, *Honey & Syrups* and *Baking Mixes* exhibit the steepest declines between the first and subsequent recommendations, suggesting that for these products, the model identifies one particularly strong pairing while additional recommendations offer diminishing returns. Conversely, categories such as *Baby Food* show a more gradual decline, indicating multiple viable recommendations with similar predicted success probabilities.

This heterogeneity highlights the potential value of a category-specific approach to determining the optimal number of recommendations to present. For categories with a steep decline, like *Baking Mixes*, a single strong recommendation might be optimal, while categories with more gradual declines could benefit from multiple recommendations that offer customers greater choice.

H.1.1 Evaluating Multiple Bundle Recommendations

Consider a focal product i with four bundle recommendations $\{(i, j_1), (i, j_2), (i, j_3), (i, j_4)\}$ where $j_1, ..., j_4$ are the top add-on products as recommended by the XGBoost model. These top products are based on the model's probability predictions for each pair, which we label p_1, p_2, p_3, p_4 respectively. Note that these probabilities represent the likelihood of purchase if each bundle were recommended in isolation. When presenting multiple recommendations simultaneously, we need to account for potential interactions between these probabilities. We provide a short and intuitive empirical demonstration that estimates the aggregate likelihood of success with multiple bundle recommendations under different scenarios.

We begin by estimating the lower bound of the probability of success. Under an optimistic view of multiple bundle recommendation, showing four bundle recommendations should perform at least as well as showing only the highest probability bundle, i.e.,

$$\text{Lower Bound} = \max\{p_1, p_2, p_3, p_4\} \tag{2}$$

In practice, this might be a conservative estimate as it assumes no incremental benefit from additional recommendations beyond the best one. This essentially replicates the analysis of Section 7 where we generalize to the entire assortment by selecting the best bundle recommendation as per the optimized policy.

Analogously, if we assume that each bundle is evaluated independently, then the probability of purchasing at least one product pair is would give us the upper bound:

$$\text{Upper Bound} = 1 - \prod_{i=1}^{4}(1 - p_i) \tag{3}$$

This represents an optimistic scenario where purchase decisions for each set of recommendations are completely independent, which is unlikely in practice since all recommendations share the same focal product.

We interpolate between these bounds using a sequential evaluation model with position effects. The model assumes customers evaluate bundles sequentially, with decreasing attention paid to lower-

ranked recommendations (Craswell et al., 2008; Dupret and Piwowarski, 2008; Ghose et al., 2014; Ursu, 2018; Gao et al., 2021; Rustichini et al., 2023). More specifically, the recommended pairs are ordered by their standalone probabilities: $p_1 \geq p_2 \geq p_3 \geq p_4$. A position effect factor $\alpha \in [0, 1]$ represents diminishing attention to lower-ranked options. Consequently, the probability of purchasing each bundled recommendation in the new model is:

$$P(\text{purchase}(i, j_1)) = p_1 \tag{4}$$

$$P(\text{purchase}(i, j_2)) = (1 - p_1) \times \alpha \times p_2 \tag{5}$$

$$P(\text{purchase}(i, j_3)) = (1 - p_1) \times (1 - \alpha \times p_2) \times \alpha^2 \times p_3 \tag{6}$$

$$P(\text{purchase}(i, j_4)) = (1 - p_1) \times (1 - \alpha \times p_2) \times (1 - \alpha^2 \times p_3) \times \alpha^3 \times p_4 \tag{7}$$

The sequential model above, while simple, has several intuitive interpretations. For instance, when $\alpha = 0$, then the model collapses to only the top bundle recommendation having a non-zero probability (equivalent to the lower bound). When $\alpha = 1$, there is no attention discounting in lower positions and the purchase probability approaches (but is not equivalent to) the upper bound. Intermediate values of α represent different scenarios where customers give lower attention to items ranked lower.

Subsequently, we can compute the expected revenue under this model as follows. Suppose r_k represents the net joint price (sum of focal product price and add-on product price minus the applied discount) for the recommended pair (i, j_k). Then,

$$\text{E[Revenue } |\alpha] = \sum_{k=1}^{4} r_k \times P(\text{purchase}(i, j_k)) \tag{8}$$

Table H18: Expected Performance of Optimized Policy with Multiple Bundle Recommendations

# Bundle Recommendations	Lower Bound	$\alpha =$ 0.05	$\alpha =$ 0.1	$\alpha =$ 0.25	Upper Bound
1	1.000	1.000	1.000	1.000	1.000
2	1.000	1.020	1.040	1.100	1.451
3	1.000	1.020	1.042	1.117	1.728
4	1.000	1.021	1.043	1.121	1.917

This table shows the expected revenue gain from showing multiple bundle recommendations for the same focal product under different assumptions of position effects as compared to showing only the best pair. The lower bound column selects only the best recommended pair, as in Section 7. The Upper Bound presents an optimistic scenario of consumers purchasing at least 1 bundle recommendation. We interpolate between the two extremes with a simple and intuitive model of sequential attention. Expected revenue measures are scaled by the value in the Lower Bound column.

To evaluate the potential impact of recommending multiple bundles, we apply this sequential evaluation model to the 16,245 focal products from Section 7.2. For each focal product, we use the XGBoost model to predict purchase probabilities for potential recommendation pairs and select the top-4 pairs based on these predictions. We then compute expected revenue under different values of α and different numbers of recommended pairs.

The results are shown in Table H18 and provide potential insights for managers. The expected revenue gain is sensitive to the position effect factor α, which captures how quickly consumer attention diminishes for lower-ranked recommendations. With modest position effects (α = 0.25), showing two bundle recommendations increases expected revenue by 10% compared to showing just one pair, while showing three or four bundle recommendations yields further incremental gains of 1.7% and 2.1% respectively. This suggests that diminishing returns set in quickly, where the marginal benefit of additional recommendations decreases substantially after the second recommendation. These findings align with research on choice overload, which indicates that excessive options can overwhelm consumers and reduce purchase likelihood (Iyengar and Lepper, 2000; Scheibehenne et al., 2010).

When position effect discounting is severe (α = 0.05), adding more recommendations beyond the first yields minimal returns of ≈2%. Conversely, with moderate discounting (α = 0.1), the gains increase to ≈4%. This parametric sensitivity highlights the importance of retailers empirically estimating position discounting rates for their specific contexts before implementing multiple recommendation strategies. The upper bound scenario, representing fully independent evaluation of recommendations, provides a theoretical maximum gain of 45% for two recommendations and up to 91% for four recommendations, though these values likely overestimate the true benefit since they ignore interactions between recommendations.

Collectively, Figure H10 and Table H18 show that retailers can enhance revenue by expanding beyond single-pair recommendations, with the optimal number varying across product categories and depending on the specific position effect factor in their context. This brief analysis relies on simplifying assumptions about consumer behavior that should be validated through controlled experiments. Future work should empirically test these predictions, particularly examining how position-based attention discounting varies across categories and customer segments, before implementing multiple recommendation strategies at scale.

H.2 Higher Order Bundle Recommendations

The previous analysis explored multiple paired recommendations for the same focal product. Here, we consider higher-order bundle recommendations, i.e., recommendations containing three or four products rather than just pairs. Unlike the multiple recommendation approach, which offers several distinct 2-product pairs for each focal product, higher-order bundle recommendations consolidate multiple products into a single, larger offering. This difference is important from both theoretical and practical perspectives. Theoretically, it allows us to consider how consumers evaluate consolidated product sets rather than independent alternative pairs. Practically, it more closely resembles the bundling strategies used by retailers such as Amazon's "Frequently Bought Together" feature, which often suggests three or more complementary products.

A key challenge in constructing higher-order bundles is selecting products that collectively offer value to consumers. Simply selecting the top-ranked products based on the XGBoost model would likely result in redundant recommendations containing similar products (e.g., multiple variations of the same product type). To address this, we propose a category-diverse greedy algorithm that prioritizes

products from different categories. For each focal product i, we first identify the top-ranked add-on products $j_1, j_2, j_3, ..., j_n$ using the XGBoost model. We initialize the bundle recommendation with the highest-ranked product pair (i, j_1) and then iteratively add the next highest-ranked product j_k only if it comes from a product category different from those already in the set. This approach ensures diversity in the recommended set while maintaining the strength of individual product relationships.

To illustrate the difference between multiple recommendations and higher-order bundles, consider a focal product "Folgers Classic Roast Ground Coffee" with the following top-ranked add-on products: 1) Coffee-mate French Vanilla Creamer, 2) Coffee-mate Hazelnut Creamer, 3) Coffee-mate Caramel Latte Creamer, 4) Splenda No Calorie Sweetener Packets, and 5) Folgers Coffee Filter Packs. Under the multiple recommendations framework, pairing coffee with different types of creamers could be an effective strategy, as it can serve customers with heterogeneous taste preferences. However, when considering higher-order bundle recommendations with multiple products, it makes little sense to pair one type of coffee with three types of creamers. Instead, a more effective approach would be to move down the ranked list to include complementary but distinct products like sweetener and coffee filters.

The category-diverse approach ensures that higher-order bundles include products that serve different consumer needs. This aims to maximize the overall utility of the bundle while avoiding redundancy. While this approach is intuitively appealing, without experimental data on higher-order bundles, we need to rely on a theoretical framework to evaluate its potential effectiveness. We turn to this evaluation next.

H.2.1 Evaluating Higher Order Bundle Recommendations

Since our field experiment directly tests only 2-product bundle recommendations, we develop a framework to extrapolate the potential impact of higher-order recommendations based on our experimental findings. An important consideration is modeling how the likelihood of purchasing a bundle recommendation changes as more products are added to it.

There are a few possible options. To estimate the probability of purchasing higher-order bundles without direct experimental evidence, we could analyze historical transaction data to identify patterns in multi-product purchases. This approach would involve building predictive models with product characteristics as independent variables and multi-product purchases as the outcome. However, such approaches would not leverage the experimental variation and would be confounded by unobserved variation in purchasing behavior. We, instead, propose to use the experimentally generated pairwise purchase probabilities while accounting for the additional complexity of larger purchase baskets, allowing us to incorporate the observed consumer response to bundle recommendations in our estimation of larger bundle recommendation scenarios.

To model the purchase probability of higher-order bundle recommendations, we propose an approach that balances the individual product relationships with the practical reality that larger bundles

(or larger product baskets more broadly) are generally harder to sell (Harlam and Lodish, 1995, cf.):

$$P^{(k)} = \gamma^{(k)} \cdot \left(\prod_{j=1}^{k} p_{ij} \right)^{(\frac{1}{k-1})} \tag{9}$$

where $P^{(k)}$ is the probability of purchasing a bundled recommendation with k products, p_{ij} is the model-predicted probability of purchasing the pair of focal product i and add-on product j, and γ is a basket complexity parameter that captures how purchase likelihood changes with the size of the recommended set.

For a 2-product bundle recommendation, Equation 9 simplifies to: $P^{(2)} = p_{ij}$, which is the probability predicted by the XGBoost model. Analogously, for a 3-product bundle recommendation, $P^{(3)} = \gamma^{(3)} \cdot (p_{ij_1} \cdot p_{ij_2})^{1/2}$, which takes the geometric mean of the individual pairwise probabilities. The geometric mean addresses a technical issue: as we add more products to the recommended set, the individual pairwise probabilities capture only part of the overall appeal. A simple product of probabilities would decrease too rapidly with recommendation size, while a simple average would not account for the multiplicative effects of compatibility. The geometric mean balances these concerns, providing a measure that accounts for the compatibility of all product pairs while maintaining a reasonable scale as bundle recommendation size increases.

The basket complexity parameter $\gamma^{(k)}$ adjusts for the change in purchase likelihood with the size of the recommended set. We define a separate parameter for each bundle size ($\gamma^{(3)}$ for 3-product bundle recommendations and $\gamma^{(4)}$ for 4-product recommendations), as the relationship may not be linear. These parameters range from 0 to 1, with lower values representing steeper declines in purchase probability as the size of the recommended set increases.

We benchmark γ using historical purchase data from the retailer. Specifically, we analyze the relative frequency of different basket sizes in historical purchases within similar product categories. Let $f(k)$ represent the frequency of k-product purchases observed in historical data. γ is then estimated as:

$$\gamma^{(k)} = \frac{f(k)}{f(2)} \tag{10}$$

where $\gamma^{(3)}$ captures the relative frequency of 3-product baskets over 2-product baskets and $\gamma^{(4)}$ captures the relative frequency of 4-product baskets over 2-product baskets. This approach anchors our extrapolation in observed consumer behavior, making the estimates more reliable than using arbitrary parameter values. Based on our analysis of historical purchase patterns, we estimate $\gamma^{(3)} \approx 0.7$ and $\gamma^{(4)} \approx 0.5$, indicating that 3-product baskets occur approximately 70% as frequently as 2-product baskets, while 4-product baskets occur approximately 50% as frequently.

Table H19: Scaled Expected Revenue for Higher Order Bundle Recommendations

# of Products	$\gamma^{(3,4)} = (0.6, 0.4)$	$\gamma^{(3,4)} = (0.7, 0.5)$	$\gamma^{(3,4)} = (0.8, 0.6)$
2	1.000	1.000	1.000
3	0.816	0.952	1.088
4	0.602	0.753	0.904

This table shows the expected revenue gain from creating higher order bundle recommendations, i.e., sets with 3 or 4 products as compared to only 2-product recommendations. The expected revenue is computed under different values of $\gamma^{(3)}$ and $\gamma^{(4)}$, the basket complexity factor from Equation 9. The expected revenue values are scaled by the revenue from 2-product bundle recommendations.

Table H19 presents the expected revenue from higher-order bundles under different scenarios, scaled relative to 2-product bundle recommendations. Under our baseline estimates from historical data ($\gamma^{(3)}$ = 0.7, $\gamma^{(4)}$ = 0.5), 3-product recommendations achieve 95% of the revenue of 2-product recommendations, while 4-product recommendations achieve 75%. This suggests that, on average, larger bundle recommendations may not improve revenue over paired offerings.

However, when we consider more optimistic scenarios where consumers show greater receptivity to larger bundles ($\gamma^{(3)}$ = 0.8, $\gamma^{(4)}$ = 0.6), 3-product recommendations outperform paired ones by 8.8%. This sensitivity analysis highlights that the effectiveness of higher-order bundles depends on the underlying basket complexity in each product context. On the other hand, when the γs are set to be more conservative ($\gamma^{(3)}$ = 0.6, $\gamma^{(4)}$ = 0.4), adding more products reduces expected revenue as compared to 2-product bundle recommendations.

It is worth noting that some of the reduction in expected revenue from larger bundle recommendations could potentially be mitigated through thoughtful interface design. For instance, Amazon's "Frequently Bought Together" feature presents customers with a 3-product bundle but allows them to easily deselect individual items, enabling those interested in all three products to purchase the complete set while accommodating customers who prefer only a subset. This flexible approach could help capture the benefits of cross-product recommendations while reducing the barrier to purchase that our model attributes to larger product baskets.

The aggregated results in Table H19 mask substantial heterogeneity across product categories, as depicted in Figure H11. For categories like *Baby Food* and *Diapering*, the expected revenue increases substantially with bundle size, even under our baseline estimates. In contrast, categories like *Cleaning Tools* show diminishing returns with larger bundles. These findings highlight the potential value of a category-specific approach to bundle size determination. The optimal bundle recommendation size likely varies by product category and depends on the strength of complementarities among products in that category. Categories where products are frequently used together (like baby products) may benefit from larger bundles, while categories with weaker complementarities may perform better with smaller bundles or multiple distinct 2-product recommendations.

It is important to emphasize that this analysis is exploratory and more focused on sizing the potential opportunity of higher-order bundles rather than providing precise estimates. Our approach relies on extrapolation from 2-product experimental data combined with historical purchase patterns, and actual

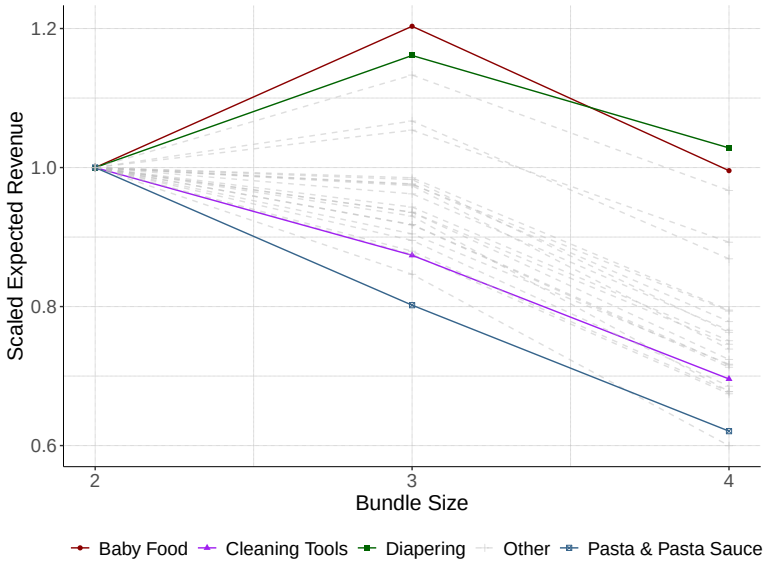


Figure H11: Heterogeneity in Expected Revenue from Higher Order Bundle Recommendations Across Product Categories

Computed with $\gamma^{(3)} = 0.7$ and $\gamma^{(4)} = 0.5$

results may vary when higher-order bundles are directly tested. The primary value of this framework is to provide directional insights into which product categories might benefit most from higher-order bundles and to estimate the potential magnitude of revenue gains under different scenarios. These insights can guide retailers in prioritizing categories for pilot testing of higher-order bundle recommendations before broader implementation. Nevertheless, these theoretical predictions should ideally be validated with experimental testing of higher-order bundles before implementation at scale.

H.3 Price Asymmetry: Focal vs. Add-on Products

An important aspect of bundle recommendations is understanding the impact of price asymmetry between the focal and add-on products. Real-world retail bundles, as in our case, are typically asymmetric, i.e., there is a "focal" product that the consumer is seeking to buy, and an "add-on" product is included with a discount to incentivize the purchase of both products. While our experiment does not incorporate price variation that would allow us to identify the optimal discount structure, our data does provide insights into how the relative prices of focal and add-on products impact bundle success.

To examine this relationship, we first test whether there is an asymmetric impact of the prices of the two products, i.e., whether consumers are more sensitive to the price of the focal product versus the add-on product, or vice versa. Using a Wald's test for the equality of the two price coefficients, we reject the null hypothesis that the coefficients are equal. Our results suggest that consumers are more sensitive to the price of the add-on product compared to the focal product. The results are shown in Table F15.

We further investigate this asymmetry by including a binary indicator in our model that takes the value 1 if the price of the focal product is greater than or equal to the price of the add-on product. The positive and significant coefficient for this indicator, as shown in Column (2) in Table F14, is consistent

with the hypothesis that bundle recommendations where the focal product's price exceeds the add-on product's price are more likely to succeed.

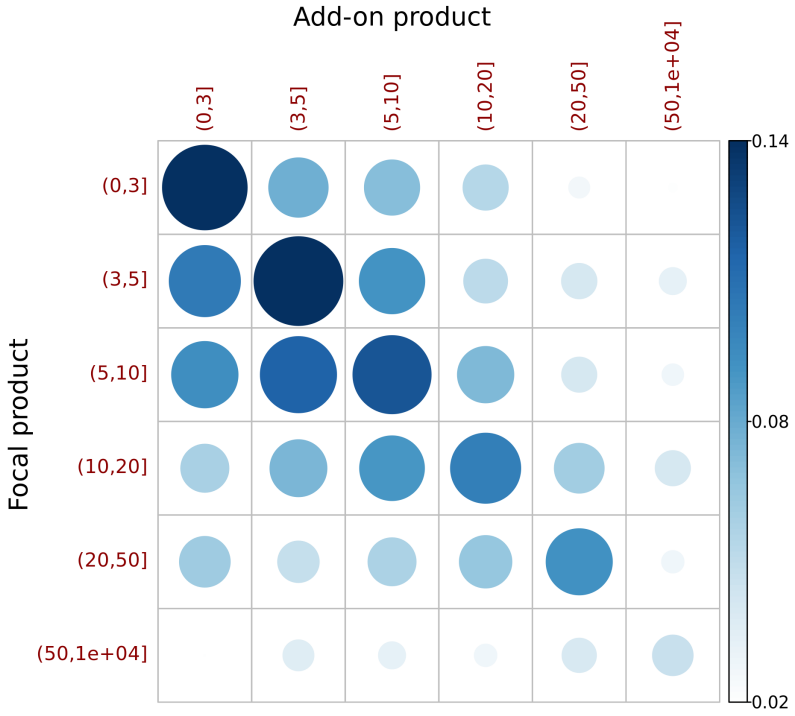


Figure H12: Predicted probabilities for bundle recommendations across price groups of focal and add-on products. Circles are sized by the number of pairs and colored by predicted probability.

We provide visual evidence of this effect in Figure H12. We create six price buckets for both focal and add-on products and then aggregate posterior predictions from the hierarchical model within these groups. Most of the mass of the predictions is in the upper left corner, where most of the bundles also lie. The visualization clearly shows the asymmetry in the effect of relative prices: consumers prefer bundles where the add-on product is cheaper than the focal product.

This price asymmetry effect can be partly explained by consumers' shopping intentions. Bundle recommendations are shown on the focal product's page, which the consumer has self-selected to view with the intention of purchasing that product. The add-on is then an additional purchase, and getting a discount on a cheaper item may be more appealing than a discount on an item of equal or higher price. Examples of successful recommended pairs where the focal product's price is strictly greater than the add-on product's price include: 1) Method 4X Laundry Detergent, Beach Sage + Method Fabric Softener, Beach Sage, 2) Blue Diamond Almonds, Bold Wasabi & Soy Sauce + Hapi Snacks Hot Wasabi Coated Green Peas, and 3) Dole California Whole Pitted Dates + Prince Of Peace 100% Natural Ginger Candy.

These findings suggest practical guidelines for retailers designing bundle recommendations: when possible, designate the higher-priced product as the focal product and offer a discount on a complementary lower-priced add-on. This structure aligns with consumers' expectations and purchase patterns, potentially increasing the success rate of bundle recommendations.

References

Bergstra, James and Yoshua Bengio, "Random Search for Hyper-Parameter Optimization," *J. Mach. Learn. Res.*, February 2012, *13* (null), 281–305.

Chen, Tianqi and Carlos Guestrin, "XGBoost: A Scala ble Tree Boosting System," in "Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining" KDD '16 ACM New York, NY, USA 2016, pp. 785–794.

Craswell, Nick, Onno Zoeter, Michael Taylor, and Bill Ramsey, "An experimental comparison of click position-bias models," in "Proceedings of the 2008 International Conference on Web Search and Data Mining" WSDM '08 Association for Computing Machinery New York, NY, USA 2008, p. 87–94.

Dupret, Georges E and Benjamin Piwowarski, "A user browsing model to predict search engine click data from past observations.," in "Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval" 2008, pp. 331–338.

Gao, Pin, Yuhang Ma, Ningyuan Chen, Guillermo Gallego, Anran Li, Paat Rusmevichientong, and Huseyin Topaloglu, "Assortment Optimization and Pricing Under the Multinomial Logit Model with Impatient Customers: Sequential Recommendation and Selection," *Operations Research*, 2021, *69* (5), 1509–1532.

Ghose, Anindya, Panagiotis G. Ipeirotis, and Beibei Li, "Examining the Impact of Ranking on Consumer Behavior and Search Engine Revenue," *Management Science*, 2014, *60* (7), 1632–1654.

Gordon, Brett R., Robert Moakler, and Florian Zettelmeyer, "Predictive Incrementality by Experimentation (PIE) for Ad Measurement," *arXiv.org*, Apr 2023.

Grbovic, Mihajlo and Haibin Cheng, "Real-Time Personalization Using Embeddings for Search Ranking at Airbnb," in "Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining" KDD '18 Association for Computing Machinery New York, NY, USA 2018, p. 311–320.

Harlam, Bari A. and Leonard M. Lodish, "Modeling Consumers' Choices of Multiple Items," *Journal of Marketing Research*, 1995, *32* (4), 404–418.

Iyengar, Sheena S and Mark R Lepper, "When Choice is Demotivating: Can One Desire too Much of a Good Thing?," *Journal of Personality and Social Psychology*, 2000, *79* (6), 995.

Kingma, Diederik P. and Jimmy Ba, "Adam: A Method for Stochastic Optimization," in Yoshua Bengio and Yann LeCun, eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Rustichini, Aldo, Philippe Domenech, Claudia Civai, and Colin G. DeYoung, "Working memory and attention in choice," *PLOS ONE*, 10 2023, *18* (10), 1–27.

Scheibehenne, Benjamin, Rainer Greifeneder, and Peter M. Todd, "Can There Ever Be Too Many Options? A Meta-Analytic Review of Choice Overload," *Journal of Consumer Research*, 02 2010, *37* (3), 409–425.

Ursu, Raluca M., "The Power of Rankings: Quantifying the Effect of Rankings on Online Consumer Search and Purchase Decisions," *Marketing Science*, 2018, *37* (4), 530–552.

Yang, Jeremy, Dean Eckles, Paramveer Dhillon, and Sinan Aral, "Targeting for Long-Term Outcomes," *Management Science*, 2024, *70* (6), 3841–3855.