

PROCEEDINGS OF THE ANALYTICS & INSIGHTS SUMMIT

(FORMERLY KNOWN AS THE SAWTOOTH SOFTWARE CONFERENCE)

October 2023

Copyright 2023

All rights reserved. No part of this volume may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from
Sawtooth Software, Inc.

FOREWORD

These proceedings are a written report of the twenty-fourth Sawtooth Software Conference, rebranded for 2023 as the Analytics & Insights Summit hosted by Sawtooth Software, held in Barcelona, Spain, May 4-5, 2023. One-hundred twenty attendees participated, both in-person and virtually.

The decision to rebrand this conference was to better represent this educational forum as the premier event for quantitative methods in marketing research. Previously called the Sawtooth Software Conference, many mistook it for a software users conference, which it was not. Relatively few speakers from Sawtooth Software have participated on the program and the general approach has been of proving and advocating research methods in an open way that is inclusive of a variety of available software platforms, both commercial and open source.

The presenters in the main research methods sessions were charged with delivering presentations of value to both the most sophisticated and least sophisticated attendees. The written versions of their presentations are contained in these Proceedings. Topics included pricing research, clustering on open-end data, experimental design, choice/conjoint analysis, MaxDiff, and market segmentation.

The papers and discussant comments are in the words of the authors and very little copyediting was performed. At the end of each of the papers are photographs of the authors and co-authors. We appreciate their cooperation for these photos! It lends a personal touch and makes it easier for readers to recognize them at the next event.

We are grateful to these authors for continuing to make this now rebranded summit such a valuable event. We feel that the Analytics & Insights Summit fulfills a multi-part mission:

- a) It advances our collective knowledge and skills,
- b) Independent authors regularly challenge the existing assumptions, research methods, and our software,
- c) It provides an opportunity for the group to renew friendships and network.

We are also especially grateful to the efforts of our steering committee who for many years now have helped this conference be such a success: Christopher Chapman, Keith Chrzan, Marco Hoogerbrugge, Joel Huber, David Lyon, Ewa Nowakowska, Bryan Orme (Chair), and Megan Peitz.

Sawtooth Software

October, 2023

CONTENTS

STRATEGIES FOR OBTAINING BUY-IN FROM STAKEHOLDERS ON CHOICE-BASED METHODS	1
<i>Patrick Meegan, Maple Street Advisors</i>	
HOW SPARSE IS TOO SPARSE? TESTING WHETHER SPARSE MAXDIFF DESIGNS WORK UNDER MORE EXTREME CONDITIONS.....	11
<i>Jon Godin, Abby Lerner, Megan Peitz, and Trevor Olsen, Numerious Inc.</i>	
REDUCED PRIMING FOR ENHANCED CONJOINT ANALYSIS.....	31
<i>Cynthia Sahm, SKIM</i>	
MONETARY OR PROPORTIONAL PRICES? A COMPARISON OF DIFFERENT APPROACHES TO SPECIFYING PRICE LEVELS IN CONJOINT ANALYSIS	41
<i>Alexandra Chirilov and James Pitcher, GfK</i>	
DESIGN AND MODELING CONSIDERATIONS WITH A DOMINATING ATTRIBUTE	59
<i>Joe Jones and Lisa Marin, Adelphi Research</i>	
COMPARING SYSTEM 1 PRIMING VS. MAXDIFF: WHICH APPROACH MEASURES BRAND PERCEPTIONS MORE ACCURATELY?	71
<i>Michael Patterson and Sonia Hundal, Radius Global Market Research</i>	
ADAPTIVE CONJOINT: TESTING BEST PRACTICES AND METHODS.....	85
<i>Zachary Levine and Kees Van der Wagt, SKIM</i>	
ALTERNATIVE-SPECIFIC CONJOINT FOR PRODUCT DEVELOPMENT AND PRICING IN TECH AND DURABLES CATEGORIES.....	95
<i>Faina Shmulyian and Tyler Dugan, Big Village</i>	
CLUSTERING OPEN-ENDED QUESTIONS: THE ALGORITHM TO AUTOMATICALLY QUANTIFY SPEECH	111
<i>Federico Adroque, KNACK Research</i>	
MANAGERIAL AND ACADEMIC CONSIDERATIONS FOR THREE APPROACHES TO WILLINGNESS TO PAY	121
<i>Bryan Orme and Keith Chrzan, Sawtooth Software; Greg Allenby, Ohio State University</i>	
SWIPE RIGHT ON SIMPLICITY: EXAMINING THE THEORETICAL AND PRACTICAL VIABILITY OF CHOICE SETS OF SIZE ONE.....	131
<i>Jeffrey P. Dotson, John Howell, and Marc Dotson, Brigham Young University; Craig Lutz, Qualtrics</i>	
FINDING CONTRASTIVE MARKET SEGMENTS WITH ARCHETYPAL ANALYSIS.....	139
<i>Jacob Nelson, Harris Poll</i>	

INTEGRATING CONSUMER GOALS IN CONJOINT USING ARCHETYPES	153
<i>Marco Vriens, Kwantum; Darin Mills and Andrew Elder, Illuminas</i>	
THE IMPACT OF MULTIPLE CLUSTER STRUCTURES ON VARIABLE SELECTION IN SEGMENTATION.....	163
<i>Joseph White, Kynetec</i>	
DESIGN AND ESTIMATION IN A CBC STUDY WITH ADDITIVE BINARY ATTRIBUTES AND PRICE	177
<i>Tommaso Gennari, Analytics with Purpose</i>	
BUILDING DESIGNS FOR INDIVIDUAL-LEVEL ESTIMATION: CONSIDERATIONS, IMPLICATIONS AND NEW TOOLS	189
<i>Megan Peitz and Trevor Olsen, Numerious</i>	
COMMENTS ON “BUILDING DESIGNS FOR INDIVIDUAL-LEVEL ESTIMATION: CONSIDERATIONS, IMPLICATIONS AND NEW TOOLS”	203
<i>Keith Chrzan, Sawtooth Software</i>	
HOW MANY ITERATIONS DO WE NEED? GUIDELINES FOR THE RIGHT NUMBER OF BURN-IN AND USED DRAWS IN HIERARCHICAL BAYES ESTIMATION.....	205
<i>Peter Kurz and Maximilian Rausch, bms - Marketing Research + Strategy</i>	

SUMMARY OF FINDINGS

The twenty-fourth Analytics & Insights Summit, hosted by Sawtooth Software (previously called the Sawtooth Software Conference) was held in Barcelona, Spain, May 4-5, 2023. The summaries below capture some of the main points of the presentations and provide a quick overview of the articles available within these Proceedings.

Strategies for Obtaining Buy-in from Stakeholders on Choice-Based Methods (Patrick Meegan, Maple Street Advisors): Patrick discussed the challenges researchers face when stakeholders resist the use of choice-based methods in survey design. He emphasized the importance of connecting the benefits of these methods to stakeholders' objectives and effectively handling stakeholder interactions to gain acceptance and alignment. Resistance can stem from various factors, such as a preference for simpler survey questions or a lack of understanding of choice-based methods. Patrick's presentation provided tactics to anticipate and address resistance, including showing visuals of the potential outcomes, creating mockups of the choice-based experience, and explaining the concepts clearly to stakeholders. He also emphasized the need to focus on outcomes, respect stakeholders' perspectives, demonstrate compromise, and avoid ego-driven responses. By following these principles and approaches, Patrick recommended that researchers can increase the likelihood of acceptance and achieve successful outcomes in their projects.

How Sparse Is Too Sparse? Testing Whether Sparse MaxDiff Designs Work under More Extreme Conditions (Jon Godin, Abby Lerner, Megan Peitz, and Trevor Olsen, Numerious): MaxDiff has become very popular and clients continue to ask researchers to push the limits. Jon and co-authors investigated best design practices for MaxDiff when facing the situation of very long descriptive text for the MaxDiff items (e.g., 250-300 characters). They tested five experimental design approaches for MaxDiff when dealing with 30 items each with 250-300 characters: Traditional MaxDiff (4 items/task, 23 tasks), Traditional Sparse MaxDiff (4 items/task, 8 tasks), Express MaxDiff (4 items/task, 12 tasks, each respondent sees a random half of the items), Extreme Sparse Pairs (2 items/task, 15 tasks), and Extreme Sparse Triplets (3 items/task, 10 tasks).

Jon and the Numerious team reported that Traditional MaxDiff (4 items/task, 23 task) did a better job of capturing individual preferences accurately but fared worse than other methods when making out-of-sample predictions and made for a more painful respondent experience with higher dropout rates, higher disqualification rates, and higher inducement to cheat while answering. On the other hand, Traditional Sparse MaxDiff (4 items/task, 8 tasks) or a best-only Paired Comparison exercise (2 items/task, 15 tasks) provided both a much better respondent experience and better out-of-sample rank-order predictions, especially when including covariates during HB utility estimation.

Reduced Priming for Enhanced Conjoint Analysis (Cynthia Sahm, SKIM): Cynthia cited findings from Kurz/Binner's 2021 Sawtooth Software Conference paper that adding 9 "behavioral calibration" questions prior to taking CBC improved the quality of the results. Although she and her colleagues at SKIM were enthusiastic to add these priming questions to their CBC studies, they received pushback from clients that the priming questions were too long.

Thus, Cynthia described how she used principal components analysis to investigate if a subset of the 9 original Kurz/Binner questions regarding brand, product innovation, and price could be simplified to a set of 3 or 4 questions.

Cynthia reported on four new CBC studies fielded using a reduced set of 4 “behavioral calibration” questions across different industries. She found that in-sample hit rates improved for 3 of 4 studies and were not harmed in the last study, validating that a subset of behavioral calibration questions is sufficient to improve model fit and worth including prior to the CBC exercise.

Monetary or Proportional Prices? A Comparison of Different Approaches to Specifying Price Levels in Conjoint Analysis (Alexandra Chirilov & James Pitcher, GfK): Alexandra and James discussed the use of conjoint analysis to address strategic pricing questions related to brand appeal and pricing power. They explored different approaches to display prices for retailers, including monetary prices with product anchoring and budget anchoring, and proportional prices. Their study compared these methods in technology and grocery retail sectors. Their research findings indicated that proportional prices, displayed as percentage deviations from expected prices, simplified the conjoint exercise setup and provided accurate results. Despite differences in price elasticities, all methods offered consistent insights into brand preference, loyalty, and switching patterns. Proportional prices emerged as a simpler and effective solution for various conjoint studies, potentially reducing project complexity and costs. Alexandra and James recommended further validation to explore its applicability in different contexts.

Design and Modeling Considerations with a Dominating Attribute (Joe Jones & Lisa Marin, Adelphi Research): Joe and Lisa investigated the effect of potentially dominating attributes (extremely important relative to others) in choice-based conjoint analysis exercises (CBC). CBC helps understand decision-making processes, often involving dominant factors like price and branding. Dominant attributes can overshadow other factors, leading to skewed results. The authors tested altering level overlap in CBC designs to reduce dominance, aiming for more accurate capture of respondent trade-offs.

Joe and Lisa used a CBC survey focusing on pain medication attributes like price, administration, safety, and side effects. They created various designs using different algorithms, testing model fit, accuracy, and attribute importance. The results suggested that designs with more level overlap, especially Sawtooth Software’s “balanced overlap”, reduced the dominance effect, leading to better model fit and increased accuracy. Alternative-specific designs with balanced overlap boosted non-dominating attribute importance, providing valuable insights for decision-making processes. The study recommended using alternative-specific designs with balanced overlap for more accurate and nuanced results in CBC analysis.

Comparing System 1 Priming vs. MaxDiff: Which Approach Measures Brand Perceptions More Accurately? (Michael Patterson & Sonia Hundal, Radius Global Market Research): Michael and Sonia compared three approaches to measuring the subconscious, System 1 (immediate, instinctual) processing: Implicit Priming Test (IPT), Emotional Valence Test (EVT), and Adaptive EVT. They examined the relationship between System 1 and System 2 measures for low and high emotional valence brands. Results showed that all three techniques

performed well, but the IPT approach stood out due to its simplicity and effectiveness. The authors also discussed the importance of incorporating System 1 processing in market research and highlights the need for a balanced perspective between System 1 and System 2. The correlations and regression models demonstrated the relationships between System 1 approaches and System 2 metrics. Overall, Michael and Sonia found that all three approaches were reliable measures of System 1 processing, but they particularly recommended the IPT approach due to its ease of use.

Adaptive Conjoint: Testing Best Practices and Methods (Zachary Levine & Kees Van der Wagt, SKIM): Traditional CBC experimental designs emphasize balance and near-orthogonality, where levels within each attribute are shown to respondents an equal number of times. Zach and Kees observed that consumers typically make decisions based on one or two key attributes rather than a holistic accounting for every possible attribute; for instance, some consumers are very brand loyal, while others may pick whichever product has the lowest price. Furthermore, consumers may focus keenly on just one or two levels within key attributes. Therefore, rather than always aiming for level balance, the authors argued that there is potential to gain greater insight and granularity for all attributes for a given consumer by aiming to emphasize each respondent's most preferred levels of each attribute.

The authors tested multiple variations of a customized adaptive CBC strategy (preference-based conjoint, or PBC) that oversamples attribute levels for later CBC tasks that are chosen by the respondent in earlier tasks. They also experimented with an approach they called “on-the-fly latent class” for further customizing the relevance of product alternatives shown to respondents in the CBC tasks.

Alternative-Specific Conjoint for Product Development and Pricing in Tech and Durables Categories (Faina Shmulyian & Tyler Dugan, Big Village): Faina and Tyler focused on the use of Choice-Based Conjoint (CBC) analysis in complex product categories, such as tech and durable goods. They compared different CBC designs, including alternative-specific design, traditional CBC, shelf test, and adaptive CBC, and examined their impact on models, simulations, conclusions, and recommendations. Their case study was digital faucets in the bathroom faucet category. They found that Alternative-specific CBC design was the most suitable for testing and optimizing innovations in this category. It allowed respondents to make meaningful trade-offs and choose from feasible combinations of attributes. The traditional CBC design simplified the structure and emphasized a smaller number of attributes. The shelf test focused on fixed configurations and price impact, while the adaptive CBC presented relevant alternatives to each respondent.

The authors highlighted the importance of selecting the appropriate design for a CBC study in complex product categories. They covered the accuracy of estimation, attribute importance scores, shares of preference, sensitivity analysis, and price sensitivity across the different designs. The alternative-specific CBC was shown to be effective in feature optimization but may have overstated price sensitivity in certain cases. Overall, they concluded that alternative-specific CBC was well-suited for complex categories and provided practical insights for researchers.

*** Clustering Open-Ended Questions: The Algorithm to Automatically Quantify Speech (Federico Adroque, KNACK Research):** Researchers often deal with open-end text and Federico demonstrated an extension of text analysis toward strategic market segmentation that goes beyond the typical word clouds or sentiment analysis. He demonstrated how open-source Python tools may be used to analyze open-ended questions efficiently and automatically in quantitative research. The algorithms he used combined descriptive statistics with machine learning to convert words into numbers, segment respondents, and generate explanatory phrases for each segment.

Federico's case study was conducted by KNACK for UNICEF in 2020, involving an open-end question posed to young students in Argentina regarding what they learned outside of school during the Covid lockdowns. His approach involved various steps, including eliminating stopwords, converting words into numbers using TF-IDF vectorizer, reducing dimensions with Principal Component Analysis (PCA), determining the optimal number of clusters, performing k-means clustering analysis, applying Non-Negative Matrix Factorization (NMF) to identify topics in each cluster, and creating correlation matrices. Future recommendations included further testing with a larger number of cases and exploring its applicability to broader open-end topics.

* Winner of Best Presentation as voted by the audience

Managerial and Academic Considerations for Three Approaches to Willingness to Pay (Bryan Orme & Keith Chrzan, Sawtooth Software, Greg Allenby, Ohio State University): Bryan, Keith, and Greg Allenby discussed three approaches to estimating respondents' willingness to pay (WTP) for features in conjoint analysis studies. The first approach was the Algebraic approach, which calculates WTP based on the price slope along with the difference in utilities between a firm's base case product and its enhanced product. The second approach was the Market Indifference approach, which involves simulating market choices and determining the price that returns the firm's enhanced product to its original share of preference. The third approach was the Social Surplus approach, which is also an algebraic approach, but it also considers the utilities of competitors and the None alternative in calculating WTP.

The authors recommended using the Market Indifference approach when the focus is on restoring market share and the Social Surplus approach when the emphasis is on restoring utility to consumers or addressing patent infringement. They illustrated results using an example dataset and explained how to applied each approach. They also discussed the typical magnitude of WTP measures and how they are affected by the strength of the firm's offering and the number of assumed competitors. Additionally, the authors examined WTP given multiple feature enhancements and provide practical considerations for implementing the different approaches.

Swipe Right on Simplicity: Examining the Theoretical and Practical Viability of Choice Sets of Size One (Jeffrey Dotson, John Howell & Marc Dotson, Brigham Young University, Craig Lutz, Qualtrics): Jeff and co-authors discussed the importance of mobile user experience in conducting surveys and studies. They proposed a solution to help overcome how challenging and unengaging conjoint analysis can be on mobile devices. The proposed solution was inspired by Tinder-style interfaces. Respondents are presented with a single product profile and asked if they would consider purchasing it by swiping right for a positive response or left for a negative

one. The authors conducted three studies to evaluate the viability of this approach and answer research questions related to data quality and user experience.

The results of Study 1 indicated that the single-alternative choice set approach provided comparable data quality and improved user experience compared to the traditional grid-based format on mobile devices. Study 2 showed that presenting approximately 80% of single alternative choice tasks can achieve equivalent statistical information to multi-alternative designs. Study 3 demonstrated that the swipeable conjoint task is quicker and more enjoyable for respondents compared to other mobile approaches. These findings suggest that implementing more intuitive and user-friendly mobile interfaces can lead to better user experience and data quality in conjoint studies. Further research is needed to explore additional aspects of this approach and address open questions related to respondent quality, parameter recovery, and boundary conditions for the swipe format. The article provides code and data related to the project for further exploration.

Finding Contrastive Market Segments with Archetypal Analysis (Jacob Nelson, Harris Poll): Jacob emphasized that market segmentation is crucial for targeted marketing. Traditional methods prioritize homogeneity within segments, missing valuable contrasts. Archetypal analysis, focusing on extremes, provides non-homogeneous yet actionable segments. It aligns with human intuition, offers flexibility in high-dimensional data, and handles outliers effectively. While not a universal solution, Jacob argued that it enriches segmentation practices, making it a valuable tool for marketers.

Integrating Consumer Goals in Conjoint Using Archetypes (Marco Vriens, Kwantum Analytics, Darin Mills & Andrew Elder, Illuminas): The authors addressed the challenges of finding impactful product feature changes in markets dominated by brand, price, and design, as well as the rapid changes in technology-driven markets. By incorporating goals and benefits, they demonstrated that the longevity of conjoint results could be extended, and more strategic insights could be obtained.

Marco, Darin, and Andrew explored various methods to integrate attributes, benefits, and goals, including laddering, benefit conjoint, hierarchical conjoint, and Archetypal analysis. The Macro conjoint focused on brand, form factor, and price, while the Micro conjoint examined health, fitness, and safety features. The analysis involved Archetypal analysis to identify switchable consumers, decision trees to predict brand switching based on goals and benefits, and regression analysis to link micro conjoint utilities to goals.

The results showed differences in brand preferences and the importance of benefits and goals in brand switching. They obtained tactical insights by identifying specific benefits and goals associated with brand loyalty. The regression analysis identified the attributes that forecasted health goals. The conclusions emphasized the humanization of conjoint analysis through the integration of goals and benefits and provided insights for strategic product roadmaps.

The Impact of Multiple Cluster Structures on Variable Selection in Segmentation (Joseph White, Kynetec): Joseph investigated the impact of multiple cluster structures on variable selection in segmentation analysis. He compared three variable selection techniques: clustvarsel (CVS), VarSelLCM (VSL), and random forests (RF). Joseph employed an

experimental design using synthetic data that varied the number of simultaneous cluster structures present in the data to assess the effectiveness of different techniques in selecting basis variables, identifying the correct number of segments, and accurately classifying records.

The results showed that CVS outperformed VSL and RF in terms of selecting effective variables, determining the right number of segments, and accurately classifying records. CVS demonstrated the ability to isolate a single cluster structure when multiple structures coexisted in the data. VSL had faster processing time but failed to remove redundancies, which negatively impacted its performance. RF retained cross-structure dimensionality, which could be explored in combination with CVS for uncovering multiple segment structures.

The findings highlighted the importance of effective variable selection in segmentation studies and its impact on uncovering true cluster structures. Joseph recommended CVS as a reliable technique for variable selection, while considering the limitations and potential benefits of other methods like RF.

Design and Estimation in a CBC Study with Additive Binary Features and Price (Tommaso Gennari, Analytics with Purpose): A well-known weakness of standard conjoint analysis modeling is when there are numerous binary (on/off) features. Main effects conjoint models typically over-predict choice of alternatives with most of the binary features “on” and underpredict choice of alternatives with most of the binary features “off”. Tommaso’s research addressed this. The goal of his project was to identify the ideal number of tiers and corresponding prices for bundled product configurations.

The conjoint design included 12 binary features presented to respondents across 10 tasks. The price structure was conditional on the number of features, with price levels varying based on feature combinations. To address the issue of diminishing returns, Tommaso introduced a predictor representing the full interaction between price and the number of features. They tested various models and ultimately chose one that showed better fit and exhibited the expected patterns of price sensitivity and diminishing returns. The market simulator built using the chosen model helped the client determine tiered bundles and their prices. Tommaso gave recommendations for designing and modeling complex conjoint studies, emphasizing the importance of collecting adequate data and exploring price sensitivity. Overall, the study proposed an alternative approach to addressing complex conjoint cases involving binary attributes, offering insights into designing and modeling such studies without requiring specialized techniques like nested logit modeling.

Building Designs for Individual-Level Estimation: Considerations, Implications and New Tools (Megan Peitz & Trevor Olsen, Numerious): Megan and Trevor explored the use of utility balanced designs in choice-based conjoint (CBC) experiments to understand consumer preferences and willingness to pay for product features. They aimed to determine if utility balanced designs could result in better predictions at the individual level compared to traditional level balanced, modest-overlap designs such as offered by Sawtooth Software’s CBC systems (both Lighthouse Studio and Discover). The study involved a CBC study regarding TVs with over 3,500 real respondents.

The results showed that utility balanced designs performed well in predicting holdout tasks featuring both utility balance and non-utility balance. Non-utility balanced designs had a harder time predicting utility-balanced choice tasks. Respondents did not seem to be fatigued by utility balanced designs. However, the authors cautioned that utility-balanced designs may result in sparse data at the interaction level. Overall, the authors suggested that utility balanced designs could be a successful strategy depending on the attributes and levels being tested, but caution should be exercised, and more research is required in this area.

How Many Iterations Do We Need? Guidelines for the Right Number of Burn-in and Used Draws in Hierarchical Bayes Estimation (Peter Kurz & Maximilian Rausch, *bms Marketing Research + Strategy*): Peter and Maximilian reported on an extensive simulation study to give practical guidelines for determining the right number of burn-in iterations, saved draws, and thinning factor in Hierarchical Bayes Multinomial Logit (HB-MNL) estimation in the context of conjoint analysis. Researchers often debate these settings, which are crucial for obtaining reliable part-worth utilities. The authors introduced a sparseness index to reflect the complexity of models and proposed guidelines based on extensive simulation across a number of CBC studies. They found that sparser models (e.g., many parameters to estimate relative to the number of choice tasks per respondent) require more iterations for convergence, and using 10,000 iterations with a thinning factor of 10 is just a good starting point (but potentially not enough) for sparse models. They provided specific recommendations based on the number of parameters in the model and the degree of sparseness. Not sparse models do well with just 20,000 burn-in followed by 10,000 used iterations. Moderately sparse models need at least 50,000 burn-in iterations, potentially more. Sparse models should use 100,000 or more burn-in iterations. They also explored the impact of long-term oscillations in HB draws, emphasizing the importance of considering these factors for accurate HB estimation in choice models.

STRATEGIES FOR OBTAINING BUY-IN FROM STAKEHOLDERS ON CHOICE-BASED METHODS

PATRICK MEEGAN¹
MAPLE STREET ADVISORS

ABSTRACT

With the proliferation of online surveys and the ease of creating surveys using a multitude of platforms, many individuals have experience creating surveys and feel comfortable in survey design. However, when researchers work with clients and propose choice-based methods that may be unfamiliar to project stakeholders, resistance isn't uncommon. Connecting the benefits of choice-based methods to stakeholders' objectives, and handling stakeholder interactions effectively will increase acceptance and expedite alignment on choice-based surveys.

Resistance can stem from a variety of sources, including a preference for simpler survey questions, an inadequate explanation of choice-based methods, or a lack of investment in stakeholder comprehension of the methods. These forms of resistance often lead to challenges in aligning the stakeholders on the researcher's recommendations. What should be done in these instances when a stakeholder wants to control how researchers obtain the necessary insights? What if they want to create their own unproven research methods? What if they see the recommended choice-based approach and reject it for being too complex, or because they don't understand it? How can a researcher anticipate resistance and get alignment with stakeholders quickly, including potential resistors?

Key Takeaways: This paper explains steps that can be taken to anticipate and diffuse objections, gain alignment with stakeholders, and identify patterns and countermeasures that will help researchers use the methods they believe are best. These tactics will assist in bringing stakeholders along, rather than "pulling rank" or expertise, which will result in better buy-in and outcomes. Researchers can take these tactics into their jobs and increase their effectiveness with stakeholders as appropriate situations arise.

INTRODUCTION

Stakeholders of research often have a lot invested in a project (time, money, reputation, career), and want to ensure they get the results and insights they need from the study. This results in the frequent involvement of a large number of people in the research design decision process to increase the likelihood of a good outcome. Most of the issues and organizational swirl observed in this process come from either unclear objectives that result in scope expansion or creep, or poorly handling the "selling" of the methods to the stakeholders.

Stakeholder expertise and perspectives are critical to improving a research plan beyond what a researcher brings as a draft. However, strong opinions can be shared by leaders and stakeholders about methods that are counterproductive to the research outcomes. At times, choice-based methods are rejected in preference for less effective research methodologies

¹ patrick.meegan@maplestreet.com

preferred by stakeholders. This can be counteracted with some deliberate planning to help stakeholders see the benefits they will receive, and how these methods deliver against better than alternatives.

Researchers who frequently use choice-based methods feel confident in their use to obtain the needed insights for a given project. However, choice-based methods are often new to stakeholders, and the lack of familiarity, apparent complexity, and level of understanding with the project team can work against a researcher. At times, some stakeholders are more skeptical in facing new approaches. With a little prep work and best practices, a researcher can shift the conversation from implying overconfidence (“I do this all the time, trust me”), to a more collaborative, justified approach.

Not all of these tactics are needed all the time, but these principles are useful in achieving successful outcomes with more efficiency, less friction, and higher confidence in the capabilities, rationale, and professionalism of the research team.

SURVEY DEVELOPMENT FLOW

Before getting into the specifics of the approach, let’s start with a survey development flow that is effective. There is a lot that could be done to expand on several of these steps. However, the scope for this paper uses steps 1–2 as a backdrop, mostly focusing on steps 3–5 and how those can be done effectively.

1. Key objective alignment with stakeholders
2. Information gathering (interviews, prior research, data sharing, etc.)
3. Draft survey including early version of choice-based example
4. Share draft with stakeholders, request feedback
5. Live workshop changes rather than via email battles
6. Get final confirmation from decision makers before fielding

SURVEY DRAFT ALIGNMENT PROCESS

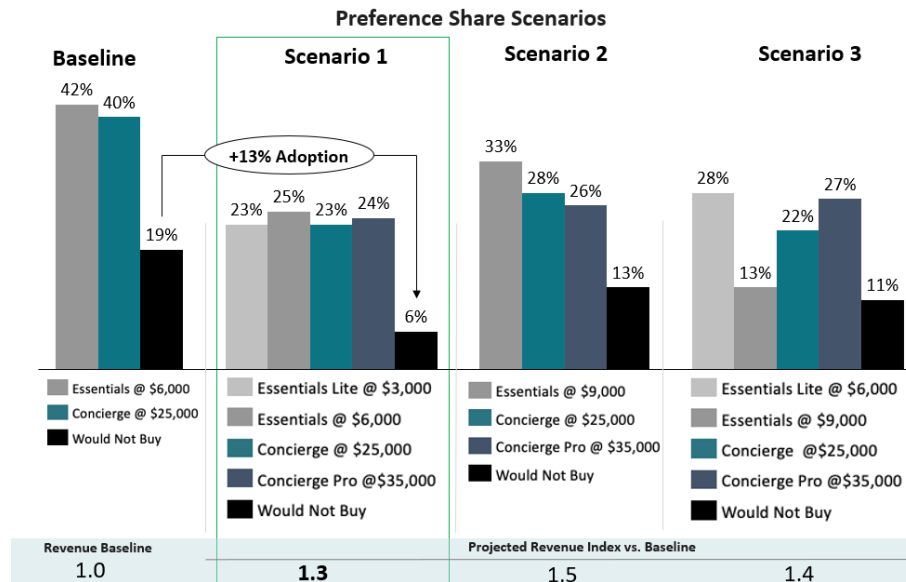
Some initial framing of the outcomes stakeholders can expect from your proposed survey can reduce friction and increase excitement for the project.

1. Show visuals of the benefits they will receive so they can wrap their minds around how this method will solve their problem and give them the info they need to make decisions. Many example visuals from prior projects (sanitized), can show the high-quality and rich information that comes from choice-based methods. Some examples could be simulations that show preference share across products/brands/features, demand curves and revenue estimations, segmentation examples, and willingness-to-pay insights. The key is not to show a dazzling array of fancy charts, but the charts or examples that are likely to help answer the question at hand *so they can see their project outcomes* in what you are showing them.

There is no perfect visual to use, but rather, using visuals to help them see the information they need coming from the methods you plan to use is the key. Walking through the examples often sparks excitement in the stakeholder group as the market research outcomes become clearer and tangible.

Show Visuals of the Potential Outcomes

1a. Simulations

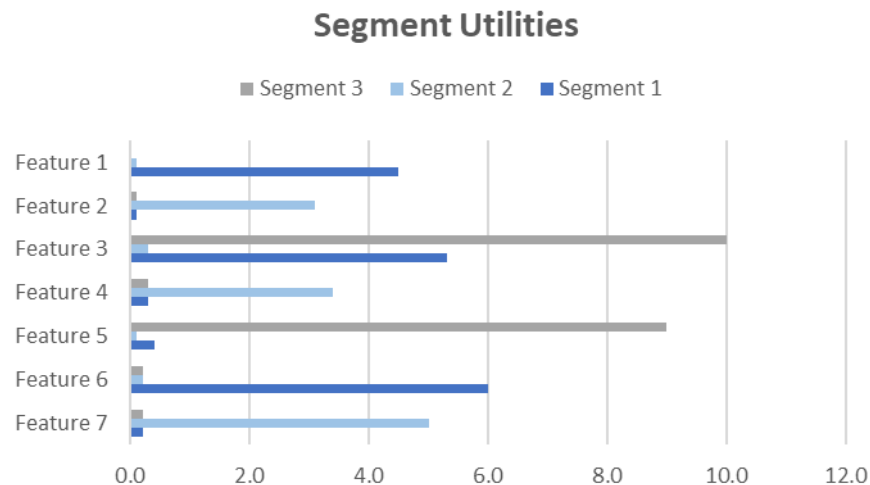


Key points that get traction:

- Better simulates a realistic market scenario than purchase likelihood scales.
- Requiring respondents to trade off reveals preferences more clearly.
- Enables dynamic simulations to answer many more questions—and to answer them in combination.
- Including price with features increases the information density and multiplies the usefulness of the data.
- Delivers willingness-to-pay insight, and can be calculated for a variety of questions and scenarios.
- In many cases, a simulation tool or spreadsheet output can be shared for future use with clients.
- More actionable than product feature-level preference information.
- Enables testing concepts in a context that includes currently available features, and features that may not exist yet/may not exist in this product yet.

- Allows for road mapping based on current and future capabilities, including combinations of features from today, more features tomorrow, and a goal state.
- Ability to optimize relative to share vs. revenue vs. profit (or adding in cost information).

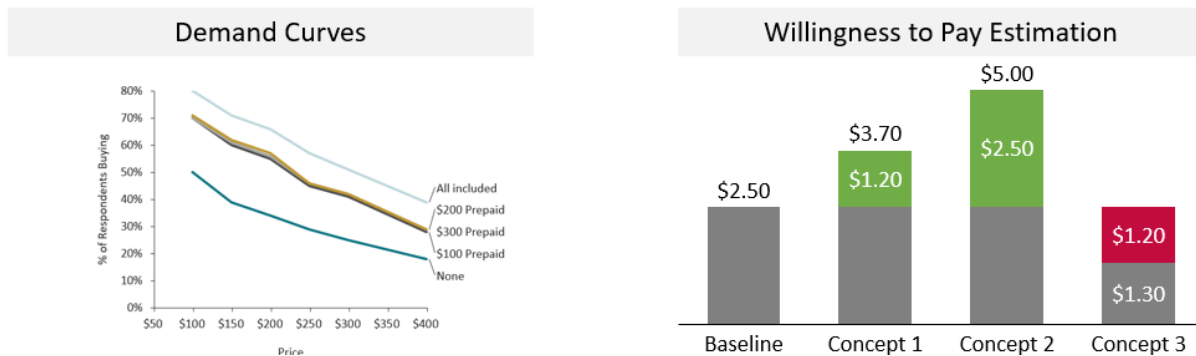
1b. Segmentation



Key points that get traction:

- Segments based on similarity of preferences/needs—more directly addresses the problem at hand.
- Does not use demographics/firmographics as the segmentation characteristics, but those variables can be used to profile the segments.
- Avoids chasing the average respondent's preferences, usually leading to “one-size-fits-none” solutions.
- Can be a collaborative process with users to discuss multiple segmentation solutions and determine the most appropriate version between researcher and client.
- Allows for segmentation to influence future product roadmaps (e.g., premium vs. budget offering) or product enhancements.

1c. Demand Curves and Willingness-to-Pay



Key points that get traction:

- Estimating change in demand from specific price points, enabling modeling of financial impacts, as well as identifying fit and revenue- maximizing price points.
 - Can be segmented by previously mentioned market segmentation data.
 - Price response data can be cut to understand the profile of respondents at varying levels of price acceptance (low, medium, high, or above \$X price, etc.).
 - Enables the calculation of willingness-to-pay differences based on feature inclusions and comparison to a variety of simulated scenarios.
2. Help them see the robustness of choice-based methods versus simple survey questions or direct methods. Examples:
 - a. “In a survey with 600 respondents, we’d have 600 responses to this question. With this choice task, at 10 conjoint tasks per respondent, we’re getting their perceptions across 6,000 choice tasks, with greater context and pricing included.”
 - b. “If we went with purchase likelihood or product preference scales, we would see that respondents want everything, and don’t have to make tradeoffs. If we force the tradeoffs, just like in real life, we can get a truer picture of the prioritization and value ascribed to specific features in the study.”
 - c. “With the greater depth of data received, we can identify preference segments that are useful in identifying attractiveness of prospects, what it takes to satisfy their needs, and which segments you want to dedicate resources to.”
 3. Create a simple mockup of the conjoint experience (for a generic exercise) with descriptions and images, and the actual choice task so they can see how it works and walk through it like a respondent. This doesn’t need to be the actual survey at hand. For example, we created a conjoint exercise using car feature selection and prices, which we share with stakeholders to illustrate how it works. Some researchers have found showing multiple screens can help comprehension, as they see features rotating. Also, others have had success showing two examples from different conjoint studies to further highlight how the process works.

4. Show your stakeholders how you'll explain the concepts in the survey so the survey respondents, like the stakeholders, can quickly go from potential overwhelm to understanding clearly what is needed—this is critical to help stakeholders know they're going to get good data from their respondents. For example, showing a conjoint explanation screen that describes the task, the actions required, the definitions for key words, and any other information that would be beneficial. Often an example image of a conjoint task with one option selected can help people to see how the process works.
5. Leverage what you know about your client's product and competitors' products to start a straw man of conjoint attributes and levels. Showing them your initial thoughts on what that could look like for them helps them wrap their minds around it. A table with the features, levels, and ranges moves the discussion along. Even if there is heavy redlining, they're working on refining the methodology proposed, rather than opting for another methodology (a win).
6. Provide hard data on how long these choice tasks take, the data quality measures you use, and the frequency with which they succeed. To show that their scenario is not the first of its kind, walk clients through the times you have worked through the nuances of a complex study like theirs.
7. Sometimes choice-based methods don't fulfill the needs of the stakeholders, indicating that other methods like interviews or focus groups will be a better fit. Be explicit about the right method to use to get the insight desired. This can be very effective in alleviating pressure on the survey to be everything to everyone.

WHEN RECEIVING RESISTANCE, FIRST UNDERSTAND THE NATURE OF THE RESISTANCE

Sometimes stakeholder resistance comes from their own perceptions or experience, but often it is due to the researcher's lack of focus on bringing them up to speed and gaining their buy-in and understanding. A non-exhaustive list of reasons why they could be resisting choice-based methods:

- They believe in another method to achieve the results
- They don't understand the choice-based methods
- Lack of familiarity with choice-based methods
- Choice-based methods were poorly explained
- They don't like choice-based methods, or find them difficult for respondents
- They are just generally skeptical
- They place higher anticipated risk on drop-off rates versus data quality, resulting in a preference for the easiest survey possible

GUIDING PRINCIPLES—RESPONDING TO RESISTANCE TO CHOICE-BASED METHODS

When resistance arises, even if you've used the practices mentioned above to increase the likelihood of acceptance of choice-based methods, it is helpful to remember these four key points:

1. Focus on the outcomes
2. Respect their perspective
3. Demonstrate compromise
4. Leave ego out of it

Along with an explanation of these points, below are some approaches to avoid, and “plays” to help researchers work through barriers.

1. **Focus on the outcomes.** The outcomes are everything. The outcomes are ultimately what they want, and they become the lynchpin for why choice-based methods are being considered at all. Go back to the survey draft practices above to show the stakeholders what they will get in terms of deliverables using choice-based methods that directly respond to the project outcomes.

Avoid: Assuming the stakeholders will just go along with what you plan to do. This assumption can set the wrong tone early in the project. It’s easy to get frustrated when you think they should just go along with it, and they think your approach is overly difficult or unnecessary.

Avoid: Going deeper than necessary on explanations or examples. Read the responses in the room and use check questions frequently so they have opportunities to clarify questions.

2. **Respect their perspectives.** Using good listening skills and brainstorming with the stakeholders in good faith will often generate better survey designs than the research team can create alone. Ask questions to understand why they have such strong opinions in certain areas; this will often reveal new options to work with.

Avoid: The temptation to be dismissive of other approaches after years of seeing the advantages of choice-based methods and the shortcomings of rating scales. Sometimes stakeholders will even invent choice-based type scenarios (they see the value of the question and outcome), but they don’t like how the choice-based method works. Focus on the outcomes of what you are doing to get stakeholders aligned.

Avoid: The battle of the pros. Sometimes stakeholders will have someone on their team with experience in survey design. Often that is very helpful, other times it can become a counterproductive tug-of-war among experts.

As Needed, Run the “Inner Circle” Play

Situation: When there is a strong opinion and/or survey expert on the team.

Play: Meet with the expert to go deeper into their concerns. Usually, their defensiveness or aggression is reduced in a smaller setting where they don’t need to look like they are still the internal expert. Give them credit for their engagement and bring them into the circle. Ask for their feedback and often you’ll gain insight that improves the research.

3. **Demonstrate compromise.** As in any negotiation, showing good faith keeps the process productive. Even if stakeholders see the researcher as the expert, it is alienating to reject

most (or all) of the feedback received from those who are often deeply knowledgeable in their business or situation.

Run the “Concession Stand” Play

Situation: The research team is receiving substantial feedback from the client stakeholders to adjust and is resisting most of them.

Play: Make concessions where there are low risks to the study outcomes—often in question phrasing or other elements of the study that you can show compromise and goodwill with the stakeholder team. It is easy to not get credit for your compromises if you don’t acknowledge them. Make sure collaborators see that you’re adjusting the survey plan according to their input. Making changes toward a choice-based method, rather than scrapping the method altogether, can keep the project moving in alignment. Get creative.

4. **Leave ego out of it.** Entrenched or emotional positions quickly become unproductive. It is not a tactic for success, even if it feels like you want to take that approach. Ultimately, you and your client want a great outcome, but you’re not agreeing on how to get it. Take the time to consider anew what could be done differently, and perhaps how your perspective is narrower than necessary. But if you must hold the line, hold it—just be sure that it’s out of serving client needs, and not ego needs.

Avoid: Falling back to a “Trust me” approach. Show them similar scenarios and how you delivered insights that meet the need/exceed the stated need.

Run the “Hold The Line” Play

Situation: Key outcomes are at risk due to stakeholder influence and changes to the survey that undermine the researcher’s ability to deliver the insights required—often because there is a choice-based research technique being rejected, questioned, or modified inappropriately.

Play: At the end of the day, it’s important to make clear that to satisfy the study objectives you can’t compromise on certain techniques that are the best way to deliver the outcomes sought. Go back to the point above about connecting the technique to the outcome and land the reasoning behind the researcher’s commitment to that approach. Play the process out to them visually—show them what you can do with your recommendation, and how the changes they are requesting limit you to deliver something that doesn’t meet project objectives.

CONCLUSION

Researchers frequently face the challenge of quickly bringing stakeholders up to speed on complex and unfamiliar approaches to obtaining insights through choice-based methods. Despite the researcher's confidence in the methods or experience using them successfully, gaining alignment with stakeholders is critical to moving projects forward and delivering the desired outcomes. Some well-tested approaches can help the research team effectively address resistance, take feedback, concede where needed, and hold the line appropriately to ultimately satisfy the stakeholder's needs for quality insights.



Patrick Meegan

HOW SPARSE IS TOO SPARSE?

TESTING WHETHER SPARSE MAXDIFF DESIGNS WORK UNDER MORE EXTREME CONDITIONS

JON GODIN
ABBY LERNER
MEGAN PEITZ
TREVOR OLSEN
NUMERIOUS INC.

EXECUTIVE SUMMARY

Clients often want to test many items using MaxDiff. Previous research shows that Sparse MaxDiff is a valid technique for testing these conditions; however, these designs typically include many alternatives per task (5 or more). What happens when the items are extremely wordy or long? Having to read through many wordy alternatives per screen across ten or twenty screens seems to be quite burdensome. But with triplets or, even worse, paired comparisons, we get much less information from each task.

In our study, we utilized a set of 30 statements about the environment, each containing between 250–300 characters, in order to ascertain which environmental issues were more or less urgent to solve now rather than leave for future generations, testing across five different design conditions: **Traditional MaxDiff** (4 statements per task, 23 tasks), **Traditional Sparse MaxDiff** (4 statements per task, 8 tasks), **Express MaxDiff** (4 statements per task, 12 tasks, 15 of 30 statements randomly selected per respondent), **Extreme Sparse Pairs** (2 items per task, 15 tasks), and **Extreme Sparse Triplets** (3 items per task, 10 tasks). We find that Traditional MaxDiff does a better job of capturing individual preferences accurately but fares worse than other methods when making out-of-sample predictions and makes for a more painful respondent experience with higher dropout rates, higher disqualification rates, and higher inducement to cheat while answering (i.e., answer randomly to finish the task more quickly). On the other hand, Traditional Sparse MaxDiff or a best-only Paired Comparison exercise provide both a much better respondent experience and better out-of-sample rank-order predictions, especially when including covariates during HB utility estimation.

BACKGROUND AND MOTIVATION

Since the time when Steve Cohen introduced Maximum Difference Scaling (MaxDiff) to the greater Sawtooth community at the 2003 Sawtooth Software Conference, MaxDiff has become a popular approach to uncovering respondent preferences among a set of items. Researchers have used MaxDiff to determine preferences for things such as advertising claims, product benefits, product messaging, images, product names, brands, features, packaging options, political voting preferences, etc.

In a typical MaxDiff exercise, respondents are shown between 2–6 items at a time, and are asked to indicate which item is best and which item is worst among the set shown (different framing can be used such as most/least motivating, most/least appealing, and others). The task is repeated many times, showing a different set of items in each task, typically using enough screens/tasks so that each item is seen by each respondent at least three times. The resulting model, using Hierarchical Bayes (HB) to estimate individual-level utilities then transforming the data into ratio-scaled probability or importance scores that sum to 100 across the items, provides the ability to understand both rank order of preference among the items as well as distances between the items (i.e., an item with a score of 10 is 2x more important or more preferable than an item with a score of 5).

As the appetite for MaxDiff grew, so did client requests to include more and more items in the set to be evaluated. With more items, many more tasks would be necessary for each item to be seen three times, but that could be burdensome for respondents. This led researchers such as Wirth and Wolfrath (2012) to test more sparse data collection methods, either using only a subset of items for each respondent (called Express MaxDiff), or still using all items, but only showing each item once to each respondent (termed Sparse MaxDiff). In a Sparse MaxDiff design, then, for 60 items you might show 15 sets of 4 items, or for 120 items, you might show 24 sets of 5 items; the important part is that each item is shown about once per respondent. Despite expectations to the contrary, Wirth and Wolfrath found that Sparse MaxDiff designs outperformed Express MaxDiff designs.

Chrzan and Peitz (2019) built upon this research with a study attempting to validate Wirth and Wolfrath's findings. They found that we can run an HB multinomial logit with fairly similar estimation when each item is shown just 1x per respondent compared to 3x per respondent (albeit with less precision at the individual level and more Bayesian smoothing).

These and other examples of prior research on Sparse MaxDiff (such as Serpetti et al., 2016) all used lists of items with relatively short statements or few total characters, such as "Is made with natural ingredients" or "Has a creamy texture." However, more and more we are being asked by our clients to test very long, high-character count statements or messages. For these exercises, do we still need to display the MaxDiff tasks in quads or quintets, or will triplets or even pairs be sufficient?

Theoretically, quads and quintets appear to provide much more information than tasks with fewer items. For example, in the example task below where we are trying to elicit color preferences, from just two clicks we are able to ascertain five preference relationships: Blue beats Red, Green, and Orange; Red beats Orange; and Green beats Orange:

Which of these colors do you most/least prefer?

(1 of 1)

Most Prefer		Least Prefer
☒	Blue	☐
☐	Green	☐
☐	Red	☐
☐	Orange	☒

The only pair we learn nothing about from this task is whether Red beats Green or Green beats Red.

However, in the following task examples with only 3 or 2 items included, we seem to gain much less information. For triplets, we learn that Blue beats Red and Green, and Red beats Green, and for pairs we only learn that Blue beats Red, but in neither case do we learn about preferences for the other colors in the design.

Most Prefer		Least Prefer
☒	Blue	☐
☐	Green	☒
☐	Red	☐

Most Prefer	
☒	Blue
☐	Red

Do we still have enough information in these smaller tasks to get stable preference estimates given that we're still only showing each item once to each respondent using a Sparse design approach?

Building on that issue, do these sparser approaches work better or worse when you have lengthy lists of long statements, such as this example which contains 296 characters:

"The average person produces 4.3 pounds of waste per day, and the U.S. accounts for 220 million tons of waste per year. This creates an environmental threat, as non-biodegradable trash gets dumped in the water, while waste from landfills generates methane, a greenhouse gas causing global warming."

With long statements, and a lot of them (30 or more), a full MaxDiff exercise showing each item to each respondent at least three times in sets of four or five items just feels very burdensome. Do we risk burning out respondents, leading to poor data quality or inducing higher dropout rates? Alternative designs—Paired Comparisons, or MaxDiff tasks shown in triplets, or

possibly even a reduced Express MaxDiff-style design where not all items are seen by each respondent—all seem like they could make things more manageable for respondents, but would the results suffer when we get less information per task? This is what we sought to find out.

CURRENT RESEARCH PLAN

In order to study the combination of high-character-count statements in a non-traditional MaxDiff exercise, we decided to focus on trying to learn people’s preferences regarding which environmental concerns they believe should be addressed now rather than pushing them off for following generations to solve.

We used a combination of old-school web searches, ChatGPT queries, and human collating and editing of these various sources into a cohesive and broad list of 30 environmental concerns, each of which ranged in length from 251 to 298 characters:

#	Statement	Num. Characters
1	Deforestation means clearing of green cover and making that land available for residential, industrial or commercial purposes. Forests cover 30% of the land, but every year tree cover is lost. Loss of forests leads to loss of biodiversity, carbon sequestration, and disruption of local communities.	298
2	Plastic pollution in oceans harms marine life and ecosystems, and can enter the human food chain. Oceans have become a giant waste dump for plastic. Unregulated disposal of waste and other materials into the ocean degrades marine and natural resources, and poses human health risks.	282
3	Water is vital for human, animal and plant survival, but water scarcity currently affects more than 40% of the world population. Growing population and industrialization are putting pressure on freshwater resources, impacting agriculture, industry, and leading to economic losses.	280
4	Air pollution in cities causes respiratory illnesses and other health problems, and contributes to climate change. Heavy metals, nitrates and plastic are among the toxins responsible for pollution, with industry and motor vehicle exhaust listed as the No. 1 pollutant.	268
5	Ocean acidity has increased in the last 250 years, but by 2100, it may shoot up by 150%. Carbon emissions are causing this impact, with 25% of total atmospheric CO2 being produced by humans. It affects ocean life and the industries that depend on it, such as fishing and tourism.	279
6	Food security around the world depends upon what condition the soil is in to produce crops. 12 million hectares of farmland is degraded each year, largely due to erosion, overgrazing, overexposure to pollutants, monoculture planting, soil compaction, and land-use conversion.	275
7	The intensive agriculture practices used to produce food have damaged the environment with the use of chemical fertilizer, pesticides and insecticides. Overuse of chemicals in agriculture harms human health too, and can lead to the development of pesticide-resistant pests.	273
8	Overfishing has a detrimental effect on natural ecosystems and leads to an imbalance of ocean life. It not only causes fishing fleets to migrate to new waters, depleting fish stocks, but also has negative effects on coastal communities that rely on fishing to support their living.	281
9	The ozone layer is an invisible layer of protection around the planet that protects life on earth from the sun’s harmful UV rays. Toxic gases are creating a hole in the ozone layer, and the depletion of this layer can lead to increased skin cancer or other health problems.	273
10	There is enough evidence to show that sea levels are rising, and the melting of Arctic ice caps and glaciers worldwide, is a major contributor. Over time, the melting of polar ice caps could lead to extensive flooding, contamination of drinking water and major changes in ecosystems.	283
11	Genetic modification of food using biotechnology is called genetic engineering. It can cause environmental problems, as an engineered gene may prove toxic to wildlife. The genetic engineering of food may also cause allergic reactions and increase resistance to antibiotics for humans.	284
12	Urban sprawl refers to population migration from high-density urban areas to low-density rural areas, causing plants and animals to be displaced from their natural environment. It leads to a decline in biodiversity, and has negative effects on the social life and economy of cities.	282
13	The average person produces 4.3 pounds of waste per day, and the U.S. accounts for 220 million tons of waste per year. This creates an environmental threat, as non-biodegradable trash gets dumped in the water, while waste from landfills generates methane, a greenhouse gas causing global warming.	296
14	People around the world use so many natural resources that we would need almost 1.5 Earths to cover all our needs. The increased use of these natural resources has led to industrialization and air pollution. Over time, natural resource depletion will lead to an energy crisis.	276

15	Human activity is leading to the extinction of species and habitats and loss of biodiversity. Ecosystems are in danger when any species' population is decimating. More than 500 species of land animals are on the brink of extinction and are likely to be lost within 20 years.	274
16	The increase of global warming from CO2 emissions is accelerating climate change. Climate change threatens the survival of millions of people, plants and animals by causing more extreme, frequent meteorological events, like droughts, fires and floods.	251
17	Illegal fishing is threatening wildlife. A shocking 640,000 tons of abandoned, lost, or otherwise discarded fishing gear is left in the world's oceans each year, which entangles and kill around 136,000 turtles, whales, seals, birds, and other sea animals.	255
18	Noise pollution is regular exposure to elevated sound levels that leads to adverse effects in humans or other living organisms. Exposure to loud noise can cause hearing loss, high blood pressure and heart disease in humans, while also negatively impacting the health and well-being of wildlife.	294
19	The world's food system is responsible for up to one third of all human-caused greenhouse gas emissions. The conventional agriculture industry has an enormous carbon footprint; it not only covers a vast amount of land, but also consumes a vast amount of freshwater.	265
20	Nuclear reactions can result in widespread contamination in air and water, aside from the loss of human life. Though nuclear reactors do not generate air pollution or carbon dioxide while operating, radioactive waste is toxic. It can cause cancer and damage to the immune system.	279
21	Cobalt is a key component of battery materials that power electric vehicles. Cobalt mining, however, is associated with many environmental and social issues. Mining regions have high radioactivity levels, and dust from pulverized rock is causing breathing problems for local communities.	287
22	Wetlands provide vital ecosystem services, such as water purification, flood control, and wildlife habitat. Wetland loss can add stress to remaining wetlands, and can also decrease habitat, landscape diversity, and connectivity among aquatic resources.	252
23	Non-native species are organisms not found naturally in an area, but are introduced as the result of human activities. Invasive non-native species are capable of causing extinctions of native plants and animals, competing with these organisms for limited resources and altering their habitats.	293
24	The world's population is more than three times larger than it was in the mid-twentieth century. Population growth not only affects food security, but also the livelihoods of farmers. As population grows, so does the demand for food, putting strain on agriculture and natural resources.	286
25	Fracking or extractive industry consists of the people, companies, and activities involved in removing oil, metals, coal, stone, and other materials from the ground. Such industry practices can cause habitat destruction, pollution, and disruption of local communities and their livelihoods.	290
26	Acid rain can be caused due to the combustion of fossil fuels or erupting volcanoes or rotting vegetation, which releases sulfur dioxide and nitrogen oxide into the atmosphere. It can also be caused by human activities. It can impact human health, wildlife, and aquatic species.	278
27	Desertification is the process by which vegetation in drylands, such as grasslands or shrublands, decreases and eventually disappears. Desertification can lead to loss of biodiversity and displacement of local communities, as well as increase the risk of zoonotic diseases.	273
28	Through the emissions from combustion of fossil-derived fuels, transportation systems contribute to degraded air quality, as well as a changing climate. Transportation also leads to noise pollution, water pollution, and affects ecosystems through multiple direct and indirect interactions.	289
29	The global demand for fashion and clothing now accounts for 10% of global carbon emissions, becoming an increasing problem. In addition to greenhouse gas emissions, textile dyeing and microplastics from various materials pollute wastewater and discarded clothing ends up in landfills.	284
30	The number of natural disasters that cost over a billion dollars has increased over the last forty years, rising from an average of 3 per year in the 1980s to 13 per year during the 2010s. Not only are natural disasters occurring more frequently, their average cost and death toll is up as well.	295

For this research, respondents would be shown only the full statements as listed above, with no use of any simplifying techniques commonly used in practice such as bolding, highlighting, or italicizing key words, providing shorter definitions on-screen with hover-overs of the full definitions, or the like. We purposefully did not want to make this easy, and perhaps sought to make it a bit painful for respondents. We think you'll agree that even reading through the statements above once is a lot to take in.

Our research design utilized five design cells, each varying either the frequency that each item would be shown to a given respondent (1x to 3x), the number of statements included per

task (2, 3, or 4), and/or the number of statements included per respondent (either a random selection of 15, or the full set of 30). The specific cells we tested were:

Cell #	MaxDiff Design Description	N Size	# Tasks
1	Traditional MaxDiff, 4 items/task, each shown 3x	302	23
2	Traditional Sparse MaxDiff, 4 items/task, each shown 1x	303	8
3	Express MaxDiff, 4 items/task, 15 items per respondent, each shown 3x	306	12
4	Extreme Sparse Pairs, 2 items/task, each shown 1x	303	15
5	Extreme Sparse Triplets, 3 items/task, each shown 1x	301	10

All designs used 200 versions. It's worth noting that for the Express MaxDiff cell, we used a design that included half of the items being shown to each respondent, which has not always been the case in earlier research, in order to give the Express approach a better chance of succeeding. Respondents in that cell would receive a randomized sampling of 15 of the 30 items, with a different randomization used for each of the 200 versions.

The study was programmed and hosted using Sawtooth Software's Lighthouse Studio. The designs for Cells 1, 2 and 3 were created within Lighthouse Studio, while the designs for Cells 4 and 5 were created using Numerious's Julia-based designer in order to ensure perfect level balance (1x) in each version of the design.

Screens for Cells 1–3 would look similar, displaying four items per task and asking respondents to indicate the most and least problematic environmental concern among the set shown. Cell 4 would only display two statements per screen, asking respondents to only indicate which statement was the most problematic, while Cell 5 would show three per screen and again ask both most and least problematic statements be identified.

In addition to the main design for each cell, we included two fixed holdouts for in-sample testing using the same structure as the main design of the cell. These holdouts were created via two-task, 1 version supplemental designs using Lighthouse Studio's MaxDiff designer. For Cell 3 holdouts (Express MaxDiff), we did not use the Serpetti et al. approach of creating an "Express Unique Anchor" holdout only showing items that a given respondent would have personally evaluated in the exercise. Therefore, the fixed holdouts for this cell are knowingly somewhat problematic, since there is no guarantee that a given respondent saw any of the four statements in each holdout during their MaxDiff exercise due to the randomization of the items entering each respondent's design.

Screenshots of each of the respondent tasks for the five cells are shown below:

Cell 1: Traditional MaxDiff

Considering these four statements, which is the **most problematic** to leave for future generations to address instead of trying to solve now? Which is **least problematic**?

(1 of 25)

Most Problematic		Least Problematic
☐	Air pollution in cities causes respiratory illnesses and other health problems, and contributes to climate change. Heavy metals, nitrates and plastic are among the toxins responsible for pollution, with industry and motor vehicle exhaust listed as the No. 1 pollutant.	☐
☐	Fracking or extractive industry consists of the people, companies, and activities involved in removing oil, metals, coal, stone, and other materials from the ground. Such industry practices can cause habitat destruction, pollution, and disruption of local communities and their livelihoods.	☐
☐	Nuclear reactions can result in widespread contamination in air and water, aside from the loss of human life. Though nuclear reactors do not generate air pollution or carbon dioxide while operating, radioactive waste is toxic. It can cause cancer and damage to the immune system.	☐
☐	Deforestation means clearing of green cover and making that land available for residential, industrial or commercial purposes. Forests cover 30% of the land, but every year tree cover is lost. Loss of forests leads to loss of biodiversity, carbon sequestration, and disruption of local communities.	☐

Cell 2: Traditional Sparse MaxDiff

Considering these four statements, which is the **most problematic** to leave for future generations to address instead of trying to solve now? Which is **least problematic**?

(1 of 10)

Most Problematic		Least Problematic
☐	There is enough evidence to show that sea levels are rising, and the melting of Arctic ice caps and glaciers worldwide, is a major contributor. Over time, the melting of polar ice caps could lead to extensive flooding, contamination of drinking water and major changes in ecosystems.	☐
☐	Fracking or extractive industry consists of the people, companies, and activities involved in removing oil, metals, coal, stone, and other materials from the ground. Such industry practices can cause habitat destruction, pollution, and disruption of local communities and their livelihoods.	☐
☐	Wetlands provide vital ecosystem services, such as water purification, flood control, and wildlife habitat. Wetland loss can add stress to remaining wetlands, and can also decrease habitat, landscape diversity, and connectivity among aquatic resources.	☐
☐	Overfishing has a detrimental effect on natural ecosystems and leads to an imbalance of ocean life. It not only causes fishing fleets to migrate to new waters, depleting fish stocks, but also has negative effects on coastal communities that rely on fishing to support their living.	☐

Cell 3: Express MaxDiff

Considering these four statements, which is the **most problematic** to leave for future generations to address instead of trying to solve now? Which is **least problematic**?

(1 of 14)

Most Problematic		Least Problematic
☐	Ocean acidity has increased in the last 250 years, but by 2100, it may shoot up by 150%. Carbon emissions are causing this impact, with 25% of total atmospheric CO2 being produced by humans. It affects ocean life and the industries that depend on it, such as fishing and tourism.	☐
☐	Overfishing has a detrimental effect on natural ecosystems and leads to an imbalance of ocean life. It not only causes fishing fleets to migrate to new waters, depleting fish stocks, but also has negative effects on coastal communities that rely on fishing to support their living.	☐
☐	Genetic modification of food using biotechnology is called genetic engineering. It can cause environmental problems, as an engineered gene may prove toxic to wildlife. The genetic engineering of food may also cause allergic reactions and increase resistance to antibiotics for humans.	☐
☐	Food security around the world depends upon what condition the soil is in to produce crops. 12 million hectares of farmland is degraded each year, largely due to erosion, overgrazing, overexposure to pollutants, monoculture planting, soil compaction, and land-use conversion.	☐

Cell 4: Sparse Paired Comparisons

Considering these two statements, which is the **most problematic** to leave for future generations to address instead of trying to solve now?

(1 of 17)

Most Problematic		Least Problematic
☐	Acid rain can be caused due to the combustion of fossil fuels or erupting volcanoes or rotting vegetation, which releases sulfur dioxide and nitrogen oxide into the atmosphere. It can also be caused by human activities. It can impact human health, wildlife, and aquatic species.	
☐	Overfishing has a detrimental effect on natural ecosystems and leads to an imbalance of ocean life. It not only causes fishing fleets to migrate to new waters, depleting fish stocks, but also has negative effects on coastal communities that rely on fishing to support their living.	

Click the "Next" button to continue...

Cell 5: Sparse Triplet MaxDiff

Considering these three statements, which is the **most problematic** to leave for future generations to address instead of trying to solve now? Which is **least problematic**?

(1 of 12)

Most Problematic		Least Problematic
☐	Plastic pollution in oceans harms marine life and ecosystems, and can enter the human food chain. Oceans have become a giant waste dump for plastic. Unregulated disposal of waste and other materials into the ocean degrades marine and natural resources, and poses human health risks.	☐
☐	Acid rain can be caused due to the combustion of fossil fuels or erupting volcanoes or rotting vegetation, which releases sulfur dioxide and nitrogen oxide into the atmosphere. It can also be caused by human activities. It can impact human health, wildlife, and aquatic species.	☐
☐	The average person produces 4.3 pounds of waste per day, and the U.S. accounts for 220 million tons of waste per year. This creates an environmental threat, as non-biodegradable trash gets dumped in the water, while waste from landfills generates methane, a greenhouse gas causing global warming.	☐

Click the "Next" button to continue...

In addition to the cell-specific in-sample holdouts, we also created two universal fixed holdout questions using a ranking task. The same two ranking tasks were shown to all respondents, regardless of design cell. Within each ranking task, we asked respondents to select a ranking for each of the four statements shown, where 1 = the most problematic issue, and 4 = the least problematic issue. An example task is shown below:

Next, here are four statements that you may or may not have seen previously. Please rank the statements in order of how problematic they are to leave for future generations to address instead of trying to solve now, where 1 is the most problematic and 4 is the least problematic.

Choose rank: ▼	Genetic modification of food using biotechnology is called genetic engineering. It can cause environmental problems, as an engineered gene may prove toxic to wildlife. The genetic engineering of food may also cause allergic reactions and increase resistance to antibiotics for humans.
Choose rank: ▼	Overfishing has a detrimental effect on natural ecosystems and leads to an imbalance of ocean life. It not only causes fishing fleets to migrate to new waters, depleting fish stocks, but also has negative effects on coastal communities that rely on fishing to support their living.
Choose rank: ▼	Desertification is the process by which vegetation in drylands, such as grasslands or shrublands, decreases and eventually disappears. Desertification can lead to loss of biodiversity and displacement of local communities, as well as increase the risk of zoonotic diseases.
Choose rank: ▼	The number of natural disasters that cost over a billion dollars has increased over the last forty years, rising from an average of 3 per year in the 1980s to 13 per year during the 2010s. Not only are natural disasters occurring more frequently, their average cost and death toll is up as well.

In all cases, the holdouts were not used in model estimation. Instead, estimated utilities from each of the design cells would be used to predict holdout choices for in-sample tasks, and item rank orders for the out-of-sample ranking tasks. The ranking questions always included four items per screen, which potentially could bias results towards those cells also showing four items per task (i.e., Cells 1–3).

Fieldwork was conducted between February 10–17, 2023, using Prodege’s peeq marketplace sample among respondents age 18+, with no other screening being used. Collected data was cleaned for speeding (< 1/3 median completion time) as well as an age mismatch (stated age asked early in survey vs. year of birth asked at the end of the survey, screening out those with a mismatch of 2 years or more). Additional data collected included gender, income, SASSY segment¹, home ownership, home area and state of residence, clean energy usage, attitudes towards climate and the environment, and political affiliation.

As an additional note on data cleaning, we did not use any on-the-fly Root Likelihood (RLH) comparisons vs. dummy respondents to clean bad cases while in field. While we like to use this approach in general practice, here we wanted to test whether any of the approaches naturally caused bad respondent behavior, so we didn’t want to screen people out prematurely. In addition, for the sparse approaches we tested, the RLH test isn’t really reliable with items being seen less than 3x per respondent, so we couldn’t apply it consistently here even if we wanted to include on-the-fly quality testing.

¹ SASSY segments were derived from the Yale Program on Climate Change Communications Six Americas Super Short Survey (SASSY), found here: <https://climatecommunication.yale.edu/visualizations-data/sassy/>

For each cell, we estimated two Hierarchical Bayes models: one with no covariates, and one including gender, income, age generation, SASSY segment, and political party affiliation as covariates. Each model utilized 20,000 burn-in and 20,000 saved iterations, otherwise using standard Lighthouse Studio estimation defaults.

Finally, we also estimated an overall model using Sawtooth Software’s stand-alone CBC/HB module by collapsing the .cho (choice) files from each of the five cells into one single file, also estimating the model twice, once without and once with the covariates listed above.

ANALYSIS OF RESULTS

In-Sample Holdouts

First, we assess in-sample validity by computing individual-level hit rates and aggregate-level Mean Absolute Errors (MAEs) when comparing actual holdout choices to those predicted from the estimated utilities for each cell. These were computed for both Best and Worst choices, but for space-saving reasons we only show the overall averages across these in the table below.

In-Sample Hit Rates (Higher is Better)

Overall Hit Rates	C1: Traditional MaxDiff	C2: Traditional Sparse MaxDiff	C3: Express MaxDiff	C4: Sparse Pairs	C5: Sparse MaxDiff Triplets
No covariates	45.9%	49.8%	46.9%	76.9%	54.2%
With covariates	46.2%	48.4%	45.7%	74.6%	53.7%
Difference	+0.3%	-1.4%	-1.2%	-2.3%	-0.5%

Both without and with covariates, Sparse Quads (Cell 2) achieve the highest hit rates among the 4-item holdout methods (Cells 1–3), and results otherwise seem reasonable. Obviously, with either pairs or triplets it’s easier to get a hit than it is with quads, as the results reflect.

For predicting individual-level choices, adding covariates to the model doesn’t seem to help and in fact for most cells slightly hurts the predictions, though the differences aren’t operationally meaningful.

In-Sample Mean Absolute Errors (Lower is Better)

Average MAEs	C1: Traditional MaxDiff	C2: Traditional Sparse MaxDiff	C3: Express MaxDiff	C4: Sparse Pairs	C5: Sparse MaxDiff Triplets
No covariates	1.7%	5.3%	3.8%	7.2%	3.5%
With covariates	1.8%	5.2%	3.5%	1.3%	4.0%
Difference	+0.1%	-0.1%	-0.3%	-5.9%	+0.5%

Moving on to MAEs, though it is sometimes the practice of academics and practitioners to tune the model exponent for each cell to minimize the within-cell MAEs, we did not take that step so the MAEs shown above are “natural.” Results-wise, we see that the full traditional MaxDiff design (Cell 1) achieves the lowest MAEs in-sample when no covariates are included in the model; Express MaxDiff also performs relatively well here given its methods bias disadvantage.

However, unlike with Hit Rates, the inclusion of covariates generally helped lower the MAEs, but only slightly in most cases, except for the Sparse Pairs cell which saw dramatic improvement. We surmise that for the Sparse Pairs the covariates are helping to reel in more extreme preferences at the individual level, leading to improved predictions of whether a given item is better than another item without overstatement.

Out-of-Sample Holdouts

Although ensuring in-sample validity is important, we feel that achieving a better ability to predict out-of-sample choices or preferences is really the gold standard for model comparisons. Here, rather than trying only to predict the overall out-of-sample rankings (considering the rankings of all 4 items at once) for each holdout ranking task, which is a very high hurdle to clear accuracy-wise, we cycled through all of the different iterations of rankings that could be derived from the data:

- **Pairs**—for any given pair in the rankings holdout, can we predict the relative ranking correctly? (18 pairs evaluated)
- **Triples**—for any given set of 3 items in the rankings holdout, can we predict the relative ranking correctly? (8 triples evaluated)
- **Quads**—for the whole set of 4 items in each ranking holdout, can we predict the relative ranking correctly? (2 quads evaluated)

In the tables that follow, we computed a weighted average across all of these splits for each cell for easier comparisons. For the Combined Model, the results represent the average across all cells. Once again, we look at both Hit Rates and MAEs for each of the methods.

Ranking Hit Rates (Higher is Better)

Ranking Hit Rates	C1: Traditional MaxDiff	C2: Traditional Sparse MaxDiff	C3: Express MaxDiff	C4: Sparse Pairs	C5: Sparse MaxDiff Triplets	Combined Model
No covariates	68.0%	63.8%	65.3%	61.0%	60.4%	64.7%
With covariates	68.2%	62.7%	61.4%	57.9%	60.2%	64.5%
Difference	+0.2%	-1.1%	-3.9%	-3.1%	-0.2%	-0.2%

As we might expect, the Sparse Cells (2, 3, and 4) perform slightly worse on hit rates than the methods where each item is shown at least 3 times to each respondent, though all methods are roughly comparable to the combined model benchmark. In this case, all cells saw ranking holdout tasks with four items each, so we wouldn't expect the Pairs or Triples to outperform the quads as we saw with the cell-specific holdouts shown earlier.

While Express MaxDiff has performed poorly in other bakeoff tests, it does surprisingly well here where we included a larger (50%) sampling of the full set of items.

As we saw for in-sample holdouts, hit rates for the rankings holdouts are generally slightly worse when covariates are included in the HB estimation.

Out-of-Sample MAEs (Lower is Better)

Ranking Aggregate MAEs	C1: Traditional MaxDiff	C2: Traditional Sparse MaxDiff	C3: Express MaxDiff	C4: Sparse Pairs	C5: Sparse MaxDiff Triplets	Combined Model
No covariates	7.2%	4.5%	11.7%	7.9%	7.9%	7.3%
With covariates	6.3%	3.1%	9.1%	4.2%	5.5%	6.5%
Pct.-point improvement	-0.9	-1.4	-2.6	-3.7	-2.4	-0.8
% Reduction in Error	-12.5%	-31.1%	-22.2%	-46.8%	-30.4%	-11.0%

For out-of-sample MAEs, Sparse Pairs and Triplets perform almost at par with Traditional MaxDiff, but Sparse Quads achieved the lowest error rate without the presence of covariates.

For the ranking holdouts, we observe marked improvement in out-of-sample predictions when using covariates across all cells; Sparse Quads (Cell 2) still perform best, but the Sparse Pairs (Cell 4) improved the most when covariates are added to the model, nearly halving the error rate achieved without covariates, and reducing the average error rate to be much closer to the overall-leading Cell 2.

Yet again, it is in the out-of-sample predictions where we continue to see Express MaxDiff suffer relative to the other methods tested.

Importance Score Comparisons

To assess the consistency of the estimated importance (probability) scores across the cells, we ran correlations of the results for each pair of test cells as well as against the overall combined model. In the table below, which shows results for the models estimated without covariates, we see that the correlations across methods are strong, with all correlations > 0.9. Cells 1 and 3 have the highest correlation with the overall model. Though Sparse Pairs (Cell 4) have the lowest correlations with other cells, they remain relatively high.

Correlations of Importance Scores by Cell

	Cell 1: Traditional MaxDiff	Cell 2: Traditional Sparse MaxDiff	Cell 3: Express MaxDiff	Cell 4: Sparse Pairs	Cell 5: Sparse MaxDiff Triplets	Combined Model
C1: Traditional MaxDiff	1.000	0.949	0.948	0.917	0.925	0.983
C2: Sparse Quads		1.000	0.938	0.915	0.911	0.969
C3: Express MaxDiff			1.000	0.903	0.961	0.978
C4: Sparse Pairs				1.000	0.902	0.949
C5: Sparse Triplets					1.000	0.966
Combined Model						1.000

Comparing the importance scores themselves across cells (again using the data from the models without covariates), we are comforted to see that the top 2 items are the same across all

cells (though the order of preference is flipped for the Sparse Pairs Cell 4), and the bottom item is the same across all cells. The relative story about the importance of the various environmental concerns is otherwise very similar across cells, with no indications of items jumping up or falling down dramatically for any cell vs. the others.

Mean Item Importance Scores

Item	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5	Combined Model
03 Water scarcity	6.44	6.73	6.65	5.18	5.54	6.02
02 Plastic pollution	5.36	5.58	5.76	5.23	5.12	5.36
16 Global warming	4.83	4.45	5.55	4.75	4.87	4.77
04 Air pollution in cities	4.30	5.06	4.92	4.74	4.83	4.69
01 Deforestation	4.57	5.20	4.83	4.74	4.18	4.56
06 Soil condition	4.03	4.72	4.92	3.95	4.73	4.35
10 Melting of Arctic ice caps and glaciers	4.36	3.54	3.71	4.09	4.39	4.05
09 Hole in the ozone layer	4.15	3.75	4.57	3.56	4.19	4.04
13 Waste	3.98	4.17	3.76	4.47	3.80	4.00
15 Extinction of species and habitats	3.91	3.69	3.26	4.40	3.67	3.76
14 Increased use of natural resources	3.80	3.71	3.42	3.78	3.62	3.76
30 Natural disasters	3.46	3.44	4.03	3.33	4.14	3.63
24 Population growth	3.68	4.36	3.72	3.00	3.59	3.62
28 Transportation system emissions	3.30	3.07	3.92	3.50	3.72	3.50
05 Ocean acidity	3.47	3.27	3.11	3.38	3.71	3.44
20 Nuclear reactions	3.08	2.95	3.70	3.33	3.48	3.26
19 Food system/agribusiness carbon footprint	3.49	3.15	3.18	2.78	2.89	3.25
07 Chemical fertilizer, pesticides, & insecticides	2.69	3.31	3.20	3.55	3.52	3.23
22 Wetland loss	3.21	2.68	2.87	3.18	2.31	2.94
27 Desertification	2.53	2.59	2.34	2.85	2.42	2.57
08 Overfishing	2.89	2.33	2.05	2.51	2.23	2.55
11 Genetic modification of food	2.42	2.59	2.48	2.76	2.08	2.47
26 Acid rain	2.34	1.94	2.54	2.32	2.94	2.46
25 Fracking or extractive industry	2.17	2.16	2.05	2.60	2.56	2.36
21 Cobalt mining	2.45	2.04	2.45	2.09	2.42	2.35
17 Abandoned fishing gear	2.33	2.55	1.74	2.19	2.36	2.25
12 Urban sprawl	2.13	2.43	1.59	2.71	1.90	2.15
23 Invasive non-native species	1.96	1.94	1.77	2.12	2.09	1.97
29 Fashion/clothing demand	1.45	1.53	1.32	2.10	1.47	1.56
18 Noise pollution	1.22	1.06	0.57	0.81	1.24	1.07

To sum up, when faced with designs that include a large number of high-character count items, sparse methods whether based on quads, pairs, or triples produce similar importance scores to traditional MaxDiff designs, but more importantly, seem to better predict out-of-sample preferences, especially when covariates are used in estimation.

RESPONDENT BEHAVIOR AND PERCEPTIONS

Beyond predictive accuracy and item preference consistency, we wanted to gauge respondent reactions to each of the designs, both behaviorally and attitudinally. First, we attempt to ascertain how burdensome each of the designs was for respondents by looking at respondent disqualification rates and perceptions of inducement to cheat during the MaxDiff exercise.

Based on the standard research DQ checks we used (less than one-third median time to complete and age mismatch), we removed significantly more respondents from Cell 1 (Traditional MaxDiff) than any of the other cells, but most particularly Cells 4 (Sparse Pairs) and 5 (Sparse Triplets), as shown in the table below:

	C1: Traditional MaxDiff	C2: Traditional Sparse MaxDiff	C3: Express MaxDiff	C4: Sparse Pairs	C5: Sparse Triplets
Removed for DQ	5.3%	3.2%	2.2%	1.6%	1.6%

The p-value of the Chi-Square statistic on disqualification rate differences across cells was 0.026, so we are confident that the DQ rate differs across the cell treatments (this exceeds the 95% threshold for the statistic to be considered statistically significant). Respondents failing DQ checks were removed prior to any subsequent analysis.

We also asked two questions *after* the MaxDiff exercise to explore whether any of the designs induced bad respondent behavior, at least from a self-reported perspective. These questions were:

1. In the hope of designing better surveys for people like you, would you please tell us . . . At any point during this exercise did you feel like selecting a random answer to get through the survey faster? Now that you are done with the exercise, it's OK to be honest and you will not be penalized for answering this honestly.
2. [If yes] You mentioned that you felt like selecting random answers in order to get through this survey faster. Did you *actually* select random answers in order to finish this survey faster? Once again, you will not be penalized for your honest answer.

Results are shown in the table below. Here again we see that, at least directionally, more respondents admitted to feeling like cheating during the exercise from Cell 1 (Traditional MaxDiff) than any other cell, and all of the four-items-per-task designs (Cells 1–3) had directionally higher rates than the tasks with only pairs or triplets (Cells 4–5). Actual admitted cheating rates are relatively comparable across tasks, ranging from ~6%–8%.

	C1: Traditional MaxDiff	C2: Traditional Sparse MaxDiff	C3: Express MaxDiff	C4: Sparse Pairs	C5: Sparse Triplets
Felt like cheating	22.2%	18.8%	16.3%	15.2%	15.6%
Admitted to cheating	6.3%	8.3%	5.9%	7.3%	7.6%

From a broader survey completion perspective, we looked at median total survey completion times as well as dropout rates. Timewise, the Sparse Triplets exercise required only 60% of the time needed for the full Traditional MaxDiff exercise. For the dropouts, we looked across cells and flagged any respondents who dropped out during the respective MaxDiff exercise they were exposed to. Here we can see that 3.2x as many dropouts occurred in Cell 1 compared to Cell 4!

	C1: Traditional MaxDiff	C2: Traditional Sparse MaxDiff	C3: Express MaxDiff	C4: Sparse Pairs	C5: Sparse Triplets
Median Survey Time	14.2	9.2	10.5	9.3	8.6
% of All Dropouts During Experiment	12.99%	9.60%	9.60%	3.95%	9.60%

So thus far we have at least directional evidence that respondents in the Traditional MaxDiff cell displayed more problematic survey behavior (higher DQ rates, higher dropout rates, and greater likelihood of feeling like cheating), and can confirm the survey length is much longer, which might induce these behaviors. How then did the respondents who completed the exercise feel about the experience?

To uncover these attitudes, we asked respondents a set of six semantic differential questions on a four-point scale, regarding their perceptions of whether the survey was:

- Long vs. Short
- Difficult vs. Easy
- Unappealing vs. Appealing
- Dull vs. Fun
- Unenjoyable vs. Enjoyable
- Confusing vs. Clear

We randomized which item was shown on the left or right as well as the order of each of the pairs during data collection. The data collected from these questions was rescaled to -4, -1, 1, 4 scaling, and items were flipped post-data collection so any “bad” items would be associated with negative scores and “good” items would be associated with positive scores. Results in the table below show that on the whole the Sparse Pairs cell (Cell 4) outperforms all of the other cells, and the Traditional MaxDiff cell (Cell 1) fares the worst by far (all results have at least directionally significant p-values from ANOVA F-tests) [*italics indicate lowest score, bold text indicates highest*]:

Semantic Differential Pair	Mean Score					ANOVA Results	
	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5	F	p-value
Long (-) vs. Short (+)	<i>-0.11</i>	0.87	0.75	0.83	0.64	10.411	<.001
Difficult (-) vs. Easy (+)	<i>1.70</i>	1.92	2.00	2.20	1.97	1.896	0.109
Unappealing (-) vs. Appealing (+)	<i>1.35</i>	1.77	1.59	1.85	1.65	2.310	0.056
Dull (-) vs. Fun (+)	<i>0.87</i>	1.32	1.16	1.63	1.28	4.800	<.001
Unenjoyable (-) vs. Enjoyable (+)	<i>1.38</i>	1.57	1.68	1.93	1.59	2.422	0.047
Confusing (-) vs. Clear (+)	2.29	2.32	2.53	2.78	2.22	3.391	0.009

Lastly, we asked respondents two open-ended questions regarding what they liked and disliked about their survey experience. NLP count vectorization of the resulting comments was conducted using Python. Several patterns emerged from the data.

In terms of likes, “easy” was mentioned ~2x more frequently by Cell 4 respondents, “nothing” was mentioned 1.5x more frequently for Cell 1 than for Cells 4 or 5, and “think” (as in “made me think”) was mentioned 1.65x more frequently for Cells 2–5 than for Cell 1.

Cell 1 Traditional MaxDiff - Likes



Cell 2 Sparse Quads - Likes



Cell 3 Express MaxDiff - Likes



Cell 4 Sparse Pairs - Likes



Cell 5 Sparse Triplets - Likes



In terms of dislikes, Cell 1 respondents used “long” 2x–4x more than other cells, “hard” was mentioned 2x less in Cell 4 than any other cell, and “repetitive” was not in the top 25 most mentioned words for Cell 4, whereas it was mentioned 2x more in Cell 1 than in Cells 2, 3, or 5.

Cell 1 Traditional MaxDiff - Dislikes



Cell 2 Sparse Quads - Dislikes



Cell 3 Express MaxDiff - Dislikes



Cell 4 Sparse Pairs - Dislikes



Cell 5 Sparse Triplets - Dislikes



So, in addition to somewhat worse respondent behavior for a Traditional MaxDiff exercise composed of many very long statements, the respondents in that cell expressed fewer positive perceptions of and stronger negative feelings toward the exercise than respondents experiencing the Sparse Pairs design, and to a somewhat lesser degree, the Traditional Sparse Quads MaxDiff design.

DISCUSSION

If we can agree that out-of-sample MAE is probably the best measuring stick, with the marked improvement when covariates are added, and factoring in respondent preference, it appears that Sparse Quads or Sparse Pairs are the best approaches for lengthy high character count MaxDiff exercises.

If you were to really need to nail individual preferences at the expense of overall accuracy, perhaps you still might consider a Traditional MaxDiff for high character count lists, but if

you're willing to sacrifice a little individual-level precision for overall market accuracy, then Sparse Quads still seems to be the gold standard of Sparse methods, though Sparse Pairs is clearly preferred by respondents and fares very well as long as relevant covariates are included in the estimation. We suspect the 3x-shown methods underperform in the aggregate due to fatigue-related issues leading to response errors, which the Sparse methods don't appear to suffer from as much.

Overall Performance Ranking Scorecard

	C1: Traditional MaxDiff	C2: Traditional Sparse MaxDiff	C3: Express MaxDiff	C4: Sparse Pairs	C5: Sparse Triplets
Hit Rates without covariates	1	3	2	4	5
Hit Rates with covariates	1	2	3	5	4
OOS MAE without covariates	2	1	5	3.5	3.5
OOS MAE with covariates	4	1	5	2	3
Respondent Preference	5	2	3	1	4
Overall	5	1	4	2	3

SUMMARY

In sum, when you need to test long lists of very wordy statements, Sparse Quads or Pairs seem best. However, all approaches we tested produced importance scores that were highly correlated across cells, which is comforting.

For designs with high character count statements, Traditional MaxDiff does a better job of capturing individual preferences accurately but fares worse than other methods OOS and offers a fairly painful respondent experience with higher dropout rates, higher disqualification rates, and higher inducement to cheat being felt by respondents. Traditional Sparse MaxDiff (showing quads) or a best-only Paired Comparison exercise (with covariates included during estimation) provide both a better respondent experience and better out-of-sample rank-order predictions.

As the list size itself increases, the Pairs method may become less viable as the number of pairs required to cover all items at least once could get quite large; traditional Sparse MaxDiff should be the go-to in that case.

Express MaxDiff fared well at capturing individual-level hit rates, so if individual-level rather than market-level inferences are your goal, it might be an option, though we suggest showing at least 50% of the items to each respondent in that case.



Jon Godin



Abby Lerner



Megan Peitz



Trevor Olsen

REFERENCES

- Chrzan, Keith and Megan Peitz (2019), “Best-Worst Scaling with Many Items,” *Journal of Choice Modeling*, Vol. 30, March 2019, pp 61–72. (See <https://www.sciencedirect.com/science/article/pii/S1755534517301355?via%3Dihub>)
- Cohen, Steven H. (2003), “Maximum Difference Scaling: Improved Measures of Importance and Preference for Segmentation.” 2003 Sawtooth Software Conference Proceedings, pp 61–74, Provo, UT.
- Orme, Bryan (2019), “Sparse, Express, Bandit, Relevant Items, Tournament, Augmented, and Anchored MaxDiff—Making Sense of All Those MaxDiffs!,” Sawtooth Software Research Paper Series (available at www.sawtoothsoftware.com/resources/technical-papers).
- Serpetti, M., Ce. Gilbert, and M. Peitz (2016), “The Researcher’s Paradox: A Further Look at the Impact of Large-Scale Choice Exercises.” 2016 Sawtooth Software Conference Proceedings, pp 147–162, Provo, UT.
- Wirth, Ralph and Annette Wolfrath (2012), “Using MaxDiff to Evaluate Very Large Sets of Items.” 2012 Sawtooth Software Conference Proceedings, Provo, UT.

REDUCED PRIMING FOR ENHANCED CONJOINT ANALYSIS

CYNTHIA SAHM
SKIM

EXECUTIVE SUMMARY

At the 2021 Sawtooth Software Conference, Peter Kurz and Stefan Binner demonstrated that across nine commercial Choice-Based Conjoint (CBC) datasets, there was an improvement in hit rates by asking a series of priming questions prior to CBC tasks. These questions focused respondents' attention on their attitudes toward brand, product innovation, and price. Kurz/Binner called them "behavioral calibration" questions.

Since then, among some clients and companies, a barrier to adoption of these questions has been the length of the priming battery, leading to hesitancy to include them. We hypothesized that these questions could be reduced and still lead to model improvement, and therefore receive greater adoption. With factor analysis, we found a reduced list of 3-4 enhanced behavioral questions was able to capture 60% of the variance observed in respondent data. Using 4 enhanced behavioral questions from the factor analysis, we gathered data for 4 projects across different industries and found that in-sample hit rates improved for 3 of 4 studies and were not harmed in the last study, validating that a subset of behavioral calibration questions is sufficient to improve model fit and worth including prior to the CBC exercise. We find these results exciting and plan to introduce the use of reduced behavioral calibration questions to more studies.

PREVIOUS RESEARCH AND MOTIVATION

Peter Kurz and Stefan Binner received the "best paper" award at the 2021 Sawtooth Software Conference for their work titled "Enhancing Conjoint Analysis with a Behavioral Framework" (Kurz and Binner, 2021). They demonstrated that the inclusion of nine questions based on principles of behavioral economics around brand, pricing, and innovation can improve conjoint models. The questions include such pairs as: "I think brands differ a lot" vs. "I think brands are more or less the same." Peter and Stefan proposed that these simple pairs statements would help respondents remember their prior shopping situations and prime them to do a more realistic job in answering CBC questions. They showed that hit rates could be significantly improved using this framework and presented results for nine different CBC studies. Even simply asking these questions improved respondent performance on the CBC tasks, without using them as covariates in model estimation.

The nine statement pairs are shown in Table 1, below, and serve the following purposes according to Kurz and Binner:

1. Help respondents remember prior shopping situations.
2. Reveal patterns of buying habits, purchase behaviors, and brand and price perceptions.
3. Help respondents create a more realistic frame of reference prior to the conjoint section.
4. Use as covariates in HB estimation and segmentation variables in product simulation.

Table 1

We would like to learn a few things about you and your general thoughts, feelings, and opinions when it comes to home upkeep, construction adhesives.
Please read each pair of statements. For each pair, please indicate whether you agree with the statement on the left or the statement on the right more, and how much more.
If both statements describe your opinion well, choose the one that best describes you. If neither seems to describe you well, choose the one that comes the closest.
Select one response for each.

	Agree Left	Agree Right	
I think that brands differ a lot	☐	☐	I think that all brands are more or less the same
I always know exactly what brand I'm going to buy before I enter the shop	☐	☐	I decide what brand I'm going to buy when I'm standing in front of the shelf
I always buy the brand I bought last time	☐	☐	I switch between different brands
I compare prices very carefully before I make a choice	☐	☐	To be honest, I compare prices only superficially
I always search for special offers first	☐	☐	Special offers are not the first thing I look out for
I always know the price of the products I buy	☐	☐	I never really know what products cost
I'm always interested in new products	☐	☐	I prefer to stick to what I know
I think that products in this category need to be improved	☐	☐	I'm completely satisfied with the products as they are
I find it easy to make the right choice for me	☐	☐	I find it very difficult to make the right choice for me

The Kurz and Binner meta-analysis across 9 different CBC projects evaluated by measuring the RMSE (Root Mean Squared Error) of predictions in comparison to actual utility values, choice shares, and, in two specific studies, real market shares. They showed increases in hit rates across all studies, see Table 2. When behavioral calibration questions were shown (column 4) vs. not shown (column 3), there was marked improvement across all studies, sometimes with an improvement of as much as 10 percentage points. The hit rates improve further with the introduction of covariates to the model. The last column (“Ensemble”) shows that when calibration questions are used one-at-a-time in HB estimations as covariates (plus again simultaneously as covariates in a single model), and then ensembled across the multiple models to make predictions, hit rates continue to improve. However, use of ensemble methods was not recommended by the authors as it did not provide enough improvement to justify using versus including only covariates.

Table 2

Table 2. Hitrate in %		Behavioral Calibration Questions			
	Chance-Rate	not shown	shown	used as covariate	Ensemble
Detergent ADW	11,11	36,50	41,60	41,90	43,20
Construction adhesives	20,00	53,90	55,30	55,60	57,10
Drops	7,69	32,40	39,10	40,20	41,90
Edible Fat	14,29	41,20	49,30	51,10	53,20
None Electric Air freshener	16,67	43,50	52,40	52,80	53,50
Hair Shampoo	7,69	30,90	32,10	33,00	33,40
Potato Chips	16,67	47,10	52,40	52,60	52,90
Laundry Detergent	7,69	31,80	36,20	37,20	37,80
Super Glue	7,69	34,20	38,70	39,10	39,60

More studies have been conducted around Enhanced Behavioral Conjoint since the Kurz/Binner paper. Last year at the 2022 Sawtooth Conference, use of behavioral calibration conducted by Orme, Godin, and Olsen (2022) showed that asking behavioral calibration questions in a MaxDiff format led to further improvement in share of preference prediction accuracy versus asking the original 9 questions.

At SKIM, a barrier to adoption for Enhanced Conjoint was the length of 9 questions. This led to hesitancy to include them in surveys and we consequently wondered if a subset could still provide benefit. To this end, we conducted factor analysis on the original 9 studies of Kurtz/Binner to reduce the list. The goal was to account for 60% of the variance observed in the data for that battery of questions. We found that between 2 and 4 factors was sufficient to explain the variance in the data, see results below for factor analyses from three studies (we feel it is not necessary to display all 9), leading us to believe that a subset of questions could be sufficient to improve conjoint model fit also.

Tables 3, 4, and 5

Table 3
Laundry

Variable	Variable Label	Factor1	Factor2	Factor3	Factor4
Q6_2	Q6_2 - I always know exactly what brand I'm going to buy before I enter the shop vs I decide what brand I'm going to buy when I'm standing in front of the shelf	0.835	-0.020	-0.074	0.024
Q6_3	Q6_3 - I always buy the brand I bought last time vs I switch between different brands	0.632	0.003	0.030	-0.181
Q6_4	Q6_4 - I compare prices very carefully before I make a choice vs To be honest, I compare prices only superficially	-0.088	0.708	-0.026	0.052
Q6_5	Q6_5 - I always search for special offers first vs Special offers are not the first thing I look out for	-0.048	0.636	0.026	0.058
Q6_6	Q6_6 - I always know the price of the products I buy vs I never really know what products cost	0.204	0.311	0.123	0.040
Q6_9	Q6_9 - I find it easy to make the right choice for me vs I find it very difficult to make the right choice for me	-0.016	0.089	0.895	0.431
Q6_7	Q6_7 - I'm always interested in new products vs I prefer to stick to what I know	-0.229	0.132	0.074	0.561
Q6_1	Q6_1 - I think that brands differ a lot (1) vs I think that all brands are more or less the same (2)	0.170	-0.034	0.060	0.245
Q6_8	Q6_8 - I think that products in this category need to be improved vs I'm completely satisfied with the products as they are	-0.030	0.031	0.074	0.205

Table 4
ConstrAdh

Variable	Variable Label	Factor1	Factor2	Factor3	Factor4
Q6_2	Q6_2 - I always know exactly what brand I'm going to buy before I enter the shop vs I decide what brand I'm going to buy when I'm standing in front of the shelf	0.674	-0.020	0.049	-0.044
Q6_3	Q6_3 - I always buy the brand I bought last time vs I switch between different brands	0.346	-0.098	-0.150	0.135
Q6_8	Q6_8 - I think that products in this category need to be improved vs I'm completely satisfied with the products as they are	0.105	0.699	0.156	0.239
Q6_9	Q6_9 - I find it easy to make the right choice for me vs I find it very difficult to make the right choice for me	0.152	-0.371	0.070	-0.013
Q6_4	Q6_4 - I compare prices very carefully before I make a choice vs To be honest, I compare prices only superficially	-0.104	0.049	0.501	0.011
Q6_6	Q6_6 - I always know the price of the products I buy vs I never really know what products cost	0.084	-0.004	0.405	-0.190
Q6_5	Q6_5 - I always search for special offers first vs Special offers are not the first thing I look out for	-0.191	-0.164	0.282	0.221
Q6_7	Q6_7 - I'm always interested in new products vs I prefer to stick to what I know	-0.042	0.116	-0.005	0.409
Q6_1	Q6_1 - I think that brands differ a lot (1) vs I think that all brands are more or less the same (2)	0.185	0.040	-0.112	0.372

Table 5
Drops

Variable	Variable Label	Factor1	Factor2
Q6_4	Q6_4 - I compare prices very carefully before I make a choice vs To be honest, I compare prices only superficially	0.799	0.537
Q6_5	Q6_5 - I always search for special offers first vs Special offers are not the first thing I look out for	0.797	0.523
Q6_7	Q6_7 - I'm always interested in new products vs I prefer to stick to what I know	0.785	0.502
Q6_8	Q6_8 - I think that products in this category need to be improved vs I'm completely satisfied with the products as they are	0.748	0.582
Q6_6	Q6_6 - I always know the price of the products I buy vs I never really know what products cost	0.728	0.607
Q6_1	Q6_1 - I think that brands differ a lot (1) vs I think that all brands are more or less the same (2)	0.673	0.635
Q6_2	Q6_2 - I always know exactly what brand I'm going to buy before I enter the shop vs I decide what brand I'm going to buy when I'm standing in front of the shelf	0.527	0.816
Q6_3	Q6_3 - I always buy the brand I bought last time vs I switch between different brands	0.566	0.786
Q6_9	Q6_9 - I find it easy to make the right choice for me vs I find it very difficult to make the right choice for me	0.648	0.671

Across the 9 studies, we consistently found a few statements that were more representative of factor groupings than others. We looked at factor scores and summed the number of times each statement pair was top-loading across factor groupings, meaning the statement was “most representative” of that factor group. Pairs that most often appeared fell along the axes of brand, price, and innovation, with a last standout being “I find it easy vs. difficult to make a choice.” See Table 6 below for final counts of statement pairs loading as most-representative for factors.

Table 6

Count	Enhanced Conjoint Statement Pair
7	Q6_2 - I always know exactly what brand I'm going to buy before I enter the shop vs I decide what brand I'm going to buy when I'm standing in front of the shelf
7	Q6_4 - I compare prices very carefully before I make a choice vs To be honest, I compare prices only superficially
7	Q6_7 - I'm always interested in new products vs I prefer to stick to what I know
5	Q6_9 - I find it easy to make the right choice for me vs I find it very difficult to make the right choice for me
2	Q6_8 - I think that products in this category need to be improved vs I'm completely satisfied with the products as they are
2	Q6_1 - I think that brands differ a lot (1) vs I think that all brands are more or less the same (2)
3	Q6_3 - I always buy the brand I bought last time vs I switch between different brands
3	Q6_5 - I always search for special offers first vs Special offers are not the first thing I look out for
2	Q6_6 - I always know the price of the products I buy vs I never really know what products cost

METHODOLOGY AND APPROACH

We used existing data for the basis of comparison with the enhanced conjoint add-on questions. We chose 4 studies representing a variety of data and industries:

1. Google FitBit Study: represents the tech space, emerging products, and complex CBC
2. Weber Pellet Grills Study: represents consumer goods and standard CBC
3. Pepsi Salty Snacks Study: a SKU conjoint setup for packaged goods
4. Chick-Fil-A Breakfast Study: menu-based conjoint with multi-select menu options

We collected data with an additional sample of 200–300 respondents per study who saw the “full priming” questions prior to the conjoint task (9 behavioral calibration pairs), and another 200–300 respondents who saw only the “reduced priming” questions prior to the task (4 behavioral calibration pairs). See Table 7 below for summary counts of each study and run. Additionally, since the original samples were much larger than our follow-ups, we did a random draw of them and reduced these models to 200–300, making the comparison between original results and follow-ups more aligned.

Table 7

Project Name	Count Original	Count Original Random Draw	Count Full Priming	Count Reduced Priming	Total Priming n	Total Atts	Total Levels
Google FitBit	1995	300	292	295	587	15	66
Weber Pellet	669	250	252	255	507	10	34
Pepsi Salty Snacks	1006	250	250	259	509	4	53
Chick-Fil-A Breakfast	600	300	300	300	600	6	36

Finally, we tweaked certain phrasing for each statement pair and project to make sure that the question was sensible around the context of the study. For example, “I decide what brand I’m going to buy when I’m standing in front of the shelf” doesn’t make sense for respondents looking at a Chick-Fil-A menu, so this statement was changed to the context of fast food and “I decide what fast food brand I’m going to buy when I’m ready to eat my meal” versus “I always know exactly what fast food brand I’m going to buy well in advance of my meal.” This update of statement pairs by study ensured that the respondents were given priming questions that were realistic to the context of the conjoint.

To determine if reduced priming improved model fit, we calculated the *in-sample hit rate* for the datasets when using task 12 as the holdout. This means that we ran all models with just 11 tasks and then used the 12th task to validate how well the model was able to predict the holdout response. We used the in-sample method because we had already collected data from the 4 studies and this allowed us to use existing sample and reduce fieldwork cost, and also speed up the timeline for insights.

We ran models for all studies with the original data, follow-up data with 4 covariates, and lastly models with follow-up data and 3 covariates. We also ran models on the follow-up data with no covariates, to see if simply introducing the paired priming questions improved the model without adding them in directly to the estimation, as Kurtz/Binner previously found.

For the 4-covariate model, each covariate represented a statement pair—so one each around brand, price, innovation, and “easy choice”—see Table 6 for reference. Another set of models was run with just 3 covariates (brand, price, innovation) and we found that hit rates between 3- and 4-covariate models were similar, indicating that future research will likely not include 4 statements but just 3. However, since all data fielded showed the 4 statement pairs to respondents, it is possible that just seeing the “is easy vs. difficult to make a choice” could have still had an impact on the model.

We ran additional KPIs beyond hit rates to further validate differences between the original models (no priming) and follow-up models (reduced-priming). The list of KPIs is below:

1. Hit Rate—the percent of time the model correctly predicts the respondent’s choice from the holdout task (so higher values=better)
2. Mean Absolute Error (MAE)—the average absolute difference between the predicted versus actual respondent probabilities (lower values=better)
3. Root Mean Squared Error (RMSE)—the square root of the MSE, which is the mean of the squared difference between the actual and predicted probabilities (lower=better)
4. Log-Likelihood—the measure of the “goodness of fit” of the model (lower=better)
5. RLH—the root likelihood across the modelled 11 tasks, the average respondent RLH across the model, a measure of how well the model “fits” the raw data
6. meanRLH—the root likelihood for the holdout tasks only

RESULTS

Models run on follow-up, reduced-priming data with 3 covariates demonstrated an average improvement in hit rates of 4% across the four studies when compared to the hit rates of the original models. This aligns well with hit rates found in the Kurtz/Binner study, which varied between 1% and 9%.

See details in Tables 8-11 below, one per study.

Table 8

model	FitBit	n	RLH	HR	MAE	RMSE	logL	meanRLH
1	Original	300	0.598231	0.613333	0.305676	0.416848	-486.862	0.212458
2	Reduced only, 3covs	295	0.610773	0.625424	0.312056	0.41144	-471.526	0.206884
3	Reduced only, no covs	295	0.594274	0.620339	0.315417	0.406962	-454.582	0.207176
	<i>improve? 2 vs 1</i>		Y	Y	N	Y	Y	N
	<i>improve? 2 vs 3</i>		Y	Y	Y	N	N	N

For the FitBit study, a variety of metrics improved when comparing model 1 (Original, no priming) versus model 2 (Reduced Priming, 3covariates), including the RLH and Hit Rate. Comparing model 2 (priming with covariates) and model 3 (priming and no covariates), the results are generally the same and the simpler model 3 with no covariates has the lower log-likelihood. Based on this, we conclude that introducing reduced priming does improve model quality, but the addition of covariates to the model is not impactful.

Table 9

model	Weber Pellet	n	RLH	HR	MAE	RMSE	logL	meanRLH
1	Original	250	0.643614	0.694	0.240715	0.351747	-330.954	0.181181
2	Reduced only, 3 covs	255	0.650633	0.692157	0.237204	0.357089	-356.979	0.173978
3	Reduced only, no covs	255	0.641367	0.696078	0.240201	0.356128	-352.03	0.172057
	<i>improve? 2 vs 1</i>		Y	N	Y	N	N	N
	<i>improve? 2 vs 3</i>		Y	N	Y	N	N	Y

The Weber Pellet Grill Study showed trivial improvement in RLH for the “reduced-priming-with-covariates” model versus both the original and “reduced with no covariates” model. Similarly, there was little difference in the hit rates, with the “reduced, no covariate” model edging out the original and “reduced-with-covariates” models just slightly. For log-likelihood, the best model was the Original. Based on results from Weber Pellet, we conclude that adding the reduced priming questions prior to the conjoint section did NOT improve the model fit, though there was also little harm done based on the similarity of the KPIS.

Table 10

model	Pepsi	n	RLH	HR	MAE	RMSE	logL	meanRLH
1	Orig, n=250	250	0.250559	0.34	0.072525	0.191196	-537.566	0.012004
2	Reduced only, 3 covs	259	0.275266	0.405405	0.071428	0.186226	-528.477	0.012011
3	Reduced only, no covs	259	0.269871	0.405405	0.068958	0.185222	-518.093	0.012785
	<i>improve? 2 vs 1</i>		Y	Y	Y	Y	Y	Y
	<i>improve? 2 vs 3</i>		Y	N	N	N	N	N

Pepsi results are displayed in the table above. The hit rate showed substantial improvement over the original model, moving from 34% to 41% with the introduction of reduced priming questions. The log-likelihood also improved, signifying that adding these items enhanced the “goodness of fit” of the model. Similar to FitBit, the “reduced-priming model with no covariates” did just as well (or better) as the “reduced priming with covariates” model.

Table 11

model	Chick-Fil-A	n	RLH	HR	MAE	RMSE	logL	meanRLH
1	Original	300	0.46	0.661667	0.134273	0.253229	-310.225	0.074642
2	Reduced only, 3 covs	259	0.4738	0.699444	0.122006	0.24235	-292.292	0.076755
3	Reduced only, no covs	259	0.491624	0.701111	0.116586	0.237114	-279.783	0.079089
	<i>improve? 2 vs 1</i>		Y	Y	Y	Y	Y	Y
	<i>improve? 2 vs 3</i>		N	N	N	N	N	N

Finally, results from Chick-Fil-A Breakfast MBC also showed significant improvement in the holdout hit rate with introduction of reduced-priming questions. The values are quite similar for the “no covariate” versus with covariate runs (both 70%) but compared to the original with no enhanced behavioral priming questions, there is a 4% improvement in model fit. There is also improvement in RLH and every other measure when comparing the “reduced priming” models to the original model. Again, there looks to be little to no benefit from adding the enhanced behavioral questions as covariates to the model because the lift comes simply from asking these questions in the first place. This is quite exciting because it makes the models faster and easier to run anyway, without dealing with the bookkeeping of covariates. (It’s worth noting that Kurz/Binner 2021 and also Godin et al. 2023 found that use of covariates helped out-of-sample predictions, considered the gold standard).

CONCLUSIONS

Results from the four studies run across our four fields of interest indicate that a subset of the Kurz/Binner “Behavioral Calibration Questions” can improve model fit, both for hit rates and a variety of other metrics. Three of four studies showed an improvement in hit rate while the last did not affect it either way. We hypothesize that there may be a seasonality effect around the one that had little to no benefit (Weber Pellet Grills) and that possibly the data collected in the fall when the original study launched represented slightly different respondent perceptions than the one with the priming question follow-ups in the spring. But this is all speculation and just our guess for why we didn’t see improvement. Another possibility is that the specific wording around the questions was not precise enough to accurately put the respondent in the shopping state of mind for the Weber Pellet study. Again, speculation.

We can conclude that the reduced “Behavioral Calibration Questions” represent a useful extension to DCM exercises. Our findings suggest that a subset set of 4 of the original 9 questions continues to help respondents to recall their most recent shopping trip in a particular category and thereby positively influence answering behavior in the ensuing conjoint model, similar to the Kurz/Binner findings of 2021. Unlike those, however, we did not see that using priming questions as covariates in the model continued to improve model fit. In our experience from these four studies, the lift came from simply introducing the questions and adding them to the model did not give further benefit.

FUTURE RESEARCH

The elephant in the room here is: What did we do with the follow-up data that showed respondents the original 9 behavioral calibration questions, and how do those models compare to the ones with only 3? We ran models on respondents with the 9 questions and found that, to our

surprise, they had worse hit rates than those of respondents who saw only 4, *and also* worse hit rates than those who saw no questions at all. We found this across all studies and after re-running repeatedly. Thus far, we have no explanation for these findings besides that something we haven't found yet is in error on the 9-question models! Another limitation to these studies was low sample size, and perhaps additional data would lead to better hit rates for the 9-question models (as well as the 3, presumably).

Another unanswered question is if we would see the same improvement in hit rates if we asked just the 3 priming questions instead of the original 4, removing the one around “it is easy for me to make a decision vs. difficult.” Because even though we ran models with just 3 covariates, it is possible that this 4th question was providing enough improvement just by its presence to justify keeping in the reduced set in the future.

Finally, these results were all based on within-sample validation tests with no additional, new respondents to confirm the generalizability of the findings with out-of-sample testing. Given, however, that the Kurz/Binner study evaluated both within-sample and out-of-sample data and both showed the same improvements in results (Table 12), we feel confident that ours will too, and plan to follow up this research with additional out-of-sample confirmation.

Table 12

RMSE	within-sample				out-of-sample			
	not shown	shown	used as covariate	Ensemble	not shown	shown	used as covariate	Ensemble
Detergent ADW	2,12	2,01	1,97	1,96	2,67	2,48	2,31	2,26
Construction adhesives	1,74	1,69	1,66	1,61	2,19	1,98	1,89	1,84
Drops	2,51	2,43	2,41	2,36	3,21	3,17	2,94	2,89
Edible Fat	2,43	2,42	2,40	2,39	3,39	3,25	3,11	3,06
None Electric Air freshener	2,72	2,62	2,58	2,57	3,94	3,37	2,89	2,81
Hair Shampoo	3,21	3,23	3,22	3,20	4,63	4,65	4,71	4,61
Potato Chips	2,16	2,05	2,01	1,96	3,12	2,93	2,73	2,67
Laundry Detergent	2,38	2,19	2,14	1,99	2,99	2,74	2,54	2,44
Super Glue	1,84	1,79	1,67	1,66	3,87	2,56	2,17	2,06



Cynthia Sahn

REFERENCES

Kurz P. and Binner S. (2021), “Enhance Conjoint with a Behavioral Framework.” 2021 Sawtooth Software Conference Proceedings, pp 91–108, Provo, UT.

Orme B., Godin J., and Olsen T. (2022), “Validation and Extension of Behavioral Calibration Questions to Improve CBC Predictions,” pp 185–200, Provo, UT.

MONETARY OR PROPORTIONAL PRICES? A COMPARISON OF DIFFERENT APPROACHES TO SPECIFYING PRICE LEVELS IN CONJOINT ANALYSIS

ALEXANDRA CHIRILOV
JAMES PITCHER
GfK

ABSTRACT

In addition to answering tactical pricing problems, such as how to optimize prices across a product portfolio, conjoint analysis can also provide answers to high-level strategic price positioning questions, such as understanding your brand's breadth of appeal and pricing power. In such studies, a major challenge is to create an appropriate set of price values to display for each brand tested in the conjoint exercise. We therefore tested three ways of displaying prices of retailer brands to respondents: monetary prices using either product anchoring or budget anchoring, and the alternative method of showing proportional prices. We demonstrate that, compared to monetary prices, proportional prices make the conjoint exercise much simpler to set-up and provides more accurate results. Proportional prices may offer a good solution to many other conjoint studies where complexity needs to be kept to a minimum, but further validation is required to assess its wider application.

MOTIVATION

Interest in pricing research has been increasing recently, largely due to changing macroeconomic factors such as high inflation. For example, mentions of price elasticity in online searches increased by 30% in 2022. There is therefore the opportunity for market researchers to capitalise on this interest and provide data-driven solutions to help clients answer key business questions relating to pricing.

Conjoint analysis is widely considered the gold standard methodology for pricing research. It is most often used to answer tactical business questions relating to price optimization, that focus on modelling the impact of price changes on specific products. For example, conjoint analysis allows us to assess the impact of price changes on demand, calculate price elasticities, and ultimately determine the optimal set of prices that maximises take-up, revenue, or profit across a product portfolio.

However, you can also use conjoint analysis to answer high-level strategic business questions relating to price positioning, such as understanding your brand's breadth of appeal and pricing power and knowing how to strengthen your brand in these areas. For example, identifying the key drivers that determine why a consumer chooses one brand over another and is willing to pay more for a particular brand.

Conjoint studies used to answer tactical business questions usually involve showing respondents a series of products and varying the associated prices of those products across each conjoint task. These price variations are usually based around the current market average price of each product, which is usually easily obtained from available databases or simple desk research. However, when attempting to answer higher-level strategic business questions, expressing an

exact price for concepts in a conjoint exercise is often more difficult. This is because you are not testing specific products but testing concepts at the overall brand level, and it is difficult to assign one price value to the current price of each brand since brands usually sell multiple products. This problem is even more notable if you are conducting pricing research in the retail sector. There is no single price value that you can show for a retailer because they sell a large range of products across multiple product categories, all with different prices.

There are a number of options to get around this problem, such as basing prices around the typical spend of the consumer in retailers in the category of interest (budget anchoring). Or we can identify a product that is typically frequently bought within the category of interest and then display prices that are relevant for that product (product anchoring). Alternatively, instead of showing prices expressed as a monetary value, we can show prices expressed as percentage deviations from what a consumer would normally expect to pay. However, there is currently little guidance on how best to implement each approach, and which approach works best. We therefore conducted a large validation study to compare the approaches and to develop a point of view on each approach.

RESEARCH DESIGN

We tested three ways of displaying prices of retailers to respondents:

1. Monetary Prices—Product Anchoring
2. Monetary Prices—Budget Anchoring
3. Proportional Prices

To do this, we fielded multiple Choice Based Conjoint (CBC) surveys, each with 400 online respondents in Germany and the UK in two distinct retailer categories: technology retailers and grocery retailers. For technology retailers we compared the monetary prices—product anchoring approach with proportional prices; and for grocery retailers the monetary prices—budget anchoring approach with proportional prices. All CBC exercises consisted of 2 attributes (retailers and price), and respondents completed 12 tasks with 8 concepts and a “none of these” option. The following sections describe in detail how we showed the monetary and proportional prices to respondents.

Monetary Prices—Product Anchoring (Used for Technology Retailers)

Prior to the conjoint exercise, we asked which type of technology products respondents were likely to buy in the future and a respondent was allocated to one of six product categories they would not reject on a least fill basis:

1. Laptop
2. Electric Toothbrush
3. Electric Kettle
4. TV
5. Vacuum Cleaner
6. Washing Machine

The respondent then saw a CBC exercise framed in the context of buying within the product category they were allocated to. For example, if the respondent was allocated to the “electric toothbrush” product category, respondents were shown a CBC exercise which asked them to “imagine you were to buy your next electric toothbrush. If these were the only available retailers, from which one would you buy given the prices shown?” (Figure 1). The prices of each retailer were calculated by taking the typical price of a standard electric toothbrush and varying the price by -20%, -10%, 0%, +10%, +20% throughout the conjoint exercise.

We told respondents to “please assume that these retailers are all equally easy to visit online or require the same travel time.” This was to ensure that their choices were not biased to their proximity to the different retailers in real life and were therefore based purely on how appealing they find each retailer and the prices that retailer offers.

Figure 1: Monetary Prices—Product Anchoring Conjoint Task (Technology Retailers)

And now try this: Imagine you were to buy your next **electric toothbrush**.

If these were the only available retailers, from which one would you buy given the prices shown?

Please assume that these retailers are equally easy to visit online or require the same travel time.

 Currys £50	 Very £61	 Ao.Com £55	 Argos £61
 Littlewoods £44	Samsung.uk, Philips.uk, Apple.uk, etc. Manufacturer store or website £66	 John Lewis £55	 Tesco £50
None of these retailers			

Monetary Prices—Budget Anchoring (Used for Grocery Retailers)

Prior to the conjoint exercise, we asked how much a respondent typically spends on their weekly grocery shopping:

1. Less than 25 Pound
2. 25–50 Pound
3. 51–75 Pound
4. 76–100 Pound
5. 101–150 Pound
6. 151–200 Pound
7. More than 200 Pound

The prices shown to the respondent were then based on the mid-point of the chosen spend band and varied by -10%, -5%, 0%, +5%, +10% throughout the conjoint exercise. This anchors the prices we show to the respondent weekly budget, ensuring the prices are always relevant to what a respondent would normally pay. Respondents were told “remember, the prices shown always apply to the same set of products in each retailer” and that to assume each retailer is equally accessible.

Figure 2: Monetary Prices—Budget Anchoring Conjoint Task (Grocery Retailers)

And now try this: Imagine **your weekly grocery shopping (food, drinks, household items, toiletries etc.)**.

If these were the only available retailers, from which one would you buy given the prices shown? **Remember, the prices shown always apply to the same set of products in each retailer.**

Please assume that these retailers are equally easy to visit online or require the same travel time.

corner shop, green grocer, butcher, etc. Local independent retailer £36	 Sainsbury's £44	 Waitrose £42	 Co-op £36
 Asda £42	 Iceland £40	 Aldi £38	Getir, Weezy, Deliveroo, Gorillas, etc. Delivery only apps £40
None of these retailers			

Proportional Prices (Used for Technology and Grocery Retailers)

Instead of showing respondents a monetary value of price, prices were shown as percentage deviations from the amount they expected to pay:

1. Cheaper (-10% below)
2. Slightly Cheaper (-5% below)
3. As much as I expect to pay
4. Slightly More Expensive (+5% above)
5. More Expensive (+10% above)

Labels such as “cheaper” and “more expensive” were shown along with the percentage values to help respondents interpret the prices more easily. The question wording was similar to the monetary prices approaches (Figures 3 and 4).

Figure 3: Proportional Prices Conjoint Task (Grocery Retailers)

And now try this: Imagine **you buy your next tech, home appliance, beauty electrical or other such product**.

If these were the only available retailers, from which one would you buy given the prices shown? **Remember, the prices shown are relative to how much you expect to pay for that product.**

Please assume that these retailers are equally easy to visit online or require the same travel time.

 John Lewis more expensive, (+10% above)	 eBay cheaper (-10% below)	 Very slightly more expensive, (+5% above)	 Amazon as much as I expect to pay
 Littlewoods slightly more expensive, (+5% above)	 Currys more expensive, (+10% above)	 Ao.Com slightly cheaper (-5% below)	 Tesco cheaper (-10% below)
None of these retailers			

Figure 4: Proportional Prices Conjoint Task (Grocery Retailers)

And now try this: Imagine **your next grocery shopping trip (food, drinks, household items, toiletries etc.)**.

If these were the only available retailers, from which one would you buy given the prices shown? **Remember, the prices shown are relative to how much you expect to pay.**

Please assume that these retailers are equally easy to visit online or require the same travel time.

corner shop, green grocer, butcher, etc. Local independent retailer cheaper (-10% below)	 Sainsbury's more expensive, (+10% above)	 Waitrose slightly more expensive, (+5% above)	 Co-op cheaper (-10% below)
 Asda slightly more expensive, (+5% above)	 Iceland as much as I expect to pay	 Aldi slightly cheaper (-5% below)	Getir, Weezy, Deliveroo, Gorillas, etc. Delivery only apps as much as I expect to pay
None of these retailers			

Quality Assessment

After completing the conjoint exercise, we asked respondents the following questions on a 5-point scale to assess the respondent experience in completing the CBC:

- Willingness to Repeat
- Ease to Answer
- Ease to Read
- Interesting
- Relevance

ANALYSIS

Utility Estimation

A separate part-worth utility was estimated for each retailer. Price was estimated as a single attribute consisting of five part-worth utilities and was constrained so lower prices have a higher utility. Utilities were estimated using Hierarchical Bayes in Choice Model R. This was so we could better automate the analysis and production of preference shares across multiple categories, countries and approaches. Conjoint shares of preference were calculated without the “none” option included, so they summed to 100%. For the monetary prices—product anchoring approach, the results were aggregated across all six product categories.

Price Elasticity Calculation

We used log-log regression to compute the price elasticity for each brand for each individual respondent. Using HB estimations, we can estimate the demand for each brand for each respondent at each price point tested, keeping all other brands at their current price. Next, we transform the data by taking the natural logarithm of price and demand. The price elasticity is the beta coefficient of the log-log regression. To calculate the market price elasticity of each brand, across the whole sample of respondents, we take a weighted average of the price elasticities across all respondents, where the weight is the share of preference.

Market Share Data

Figures for real-world market shares for the different retailers were obtained from GfK’s Point of Sale (POS) panel for Technology retailers and from GfK’s Consumer Panel for Grocery retailers.

Data Quality Checks

We calculated the average time respondents took to complete each conjoint exercise, the proportion of respondents who dropped out of the survey and identified “bad” respondents as those who consistently chose the same concept (excluding the “none” option) in over 75% of tasks or exhibited different selection behaviours between the first and second halves of the conjoint exercise. Internal model validity was assessed using a holdout task.

RESULTS

Data Quality

Looking at the results of the user experience questions, we see no significant differences in respondent experience between the different approaches (Figure 5). In all approaches, the overall experience was good but there is some room to increase how relevant and interesting the exercise is.

Figure 5: Results of the User Experience Questions



The total time to complete the conjoint exercise and the proportion of respondents dropping out of the survey were consistent across the approaches and similar to what we typically see in other studies (Figure 6). However, the “none” option was chosen more often, and the proportion of “bad” respondents is much higher for monetary prices with product anchoring compared to the other approaches.

Figure 6: Results of Various Data Quality Checks

	% None	% Bad Respondents*	% Drop-Outs**	Total Time (s)
Monetary Prices – Product Anchoring	9%	14.4%	20%	102
Monetary Prices – Budget Anchoring	4%	0.4%	19%	116
Proportional Prices	6%	2.8%	20%	120

All approaches have exceptionally high internal validity (Figure 7). The model fit (RLH) is high for all approaches but is slightly higher for monetary prices with product anchoring and slightly lower for proportional prices. The hit rate is slightly lower for proportional prices than the other two approaches whereas the Mean Absolute Error (MAE) is slightly higher for monetary prices with product anchoring. No method is performing notably better than the others.

Figure 7: Results of Internal Validity: Model Fit and Hold-Out Analysis

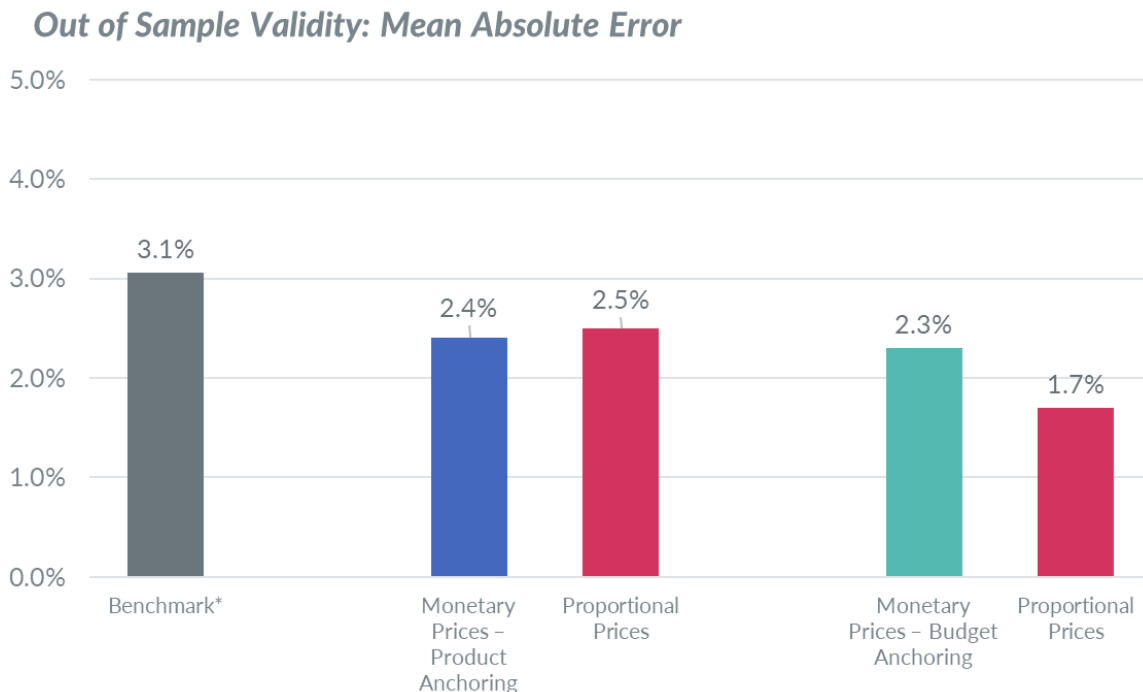
	RLH	Hit Rate (vs. holdout task)	MAE (vs. holdout task)
Monetary Prices – Product Anchoring	0.69	81.7%	1.1%
Monetary Prices – Budget Anchoring	0.65	82.8%	0.7%
Proportional Prices	0.60	77.6%	0.8%

Random Chance: 11%

Shares of Preference vs. Real-World Market Shares

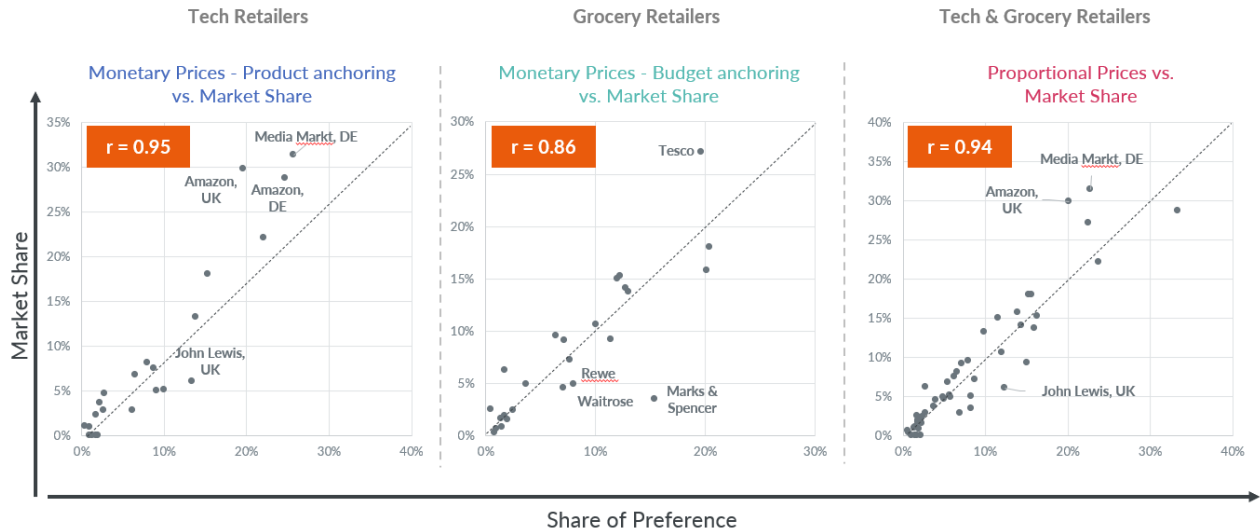
When we compare the shares of preference derived from each approach versus real-world market shares, we see that the two are closely aligned with low Mean Absolute Errors (Figure 8). The approaches outperform the benchmark value, derived from various brand-price conjoint analyses conducted for manufacturers' brands across multiple sectors. For technology retailers, both monetary and proportional prices have similar errors. While for grocery retailers, the proportional prices approach has a lower error than monetary prices.

Figure 8: Mean Absolute Error Between Conjoint Shares of Preference and Market Shares



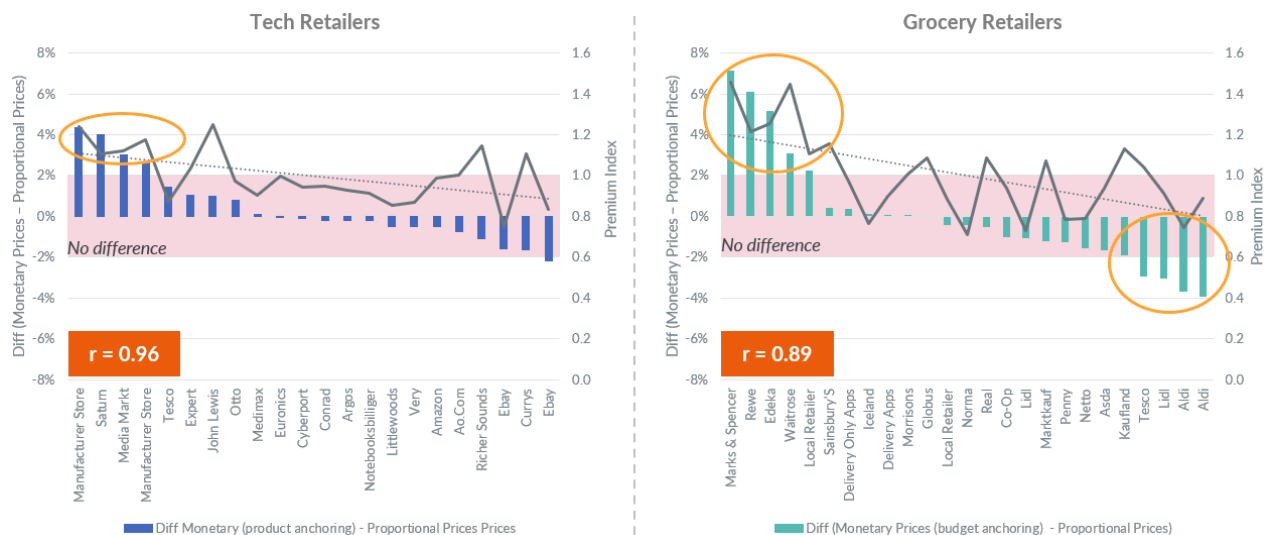
When we look at how the shares of preference for brands correlate with their market shares, we see a similar story. All approaches demonstrate a high correlation between their shares of preference and market shares (Figure 9). However, the correlation for monetary prices with budget anchoring for grocery retailers is lower than the other approaches.

Figure 9: Correlation Between Conjoint Shares of Preference and Market Shares



When we observe the relationship between differences in the shares of preference and market share and how premium the brand is perceived to be (based on stated survey question), we see that monetary approaches tend to overstate the shares of premium retailers and underestimate the share of low-cost retailers, particular for grocery retailers (Figure 10).

Figure 10: Relationship Between Differences in the Shares of Preference and Market Share and How Premium the Brand is Perceived to Be

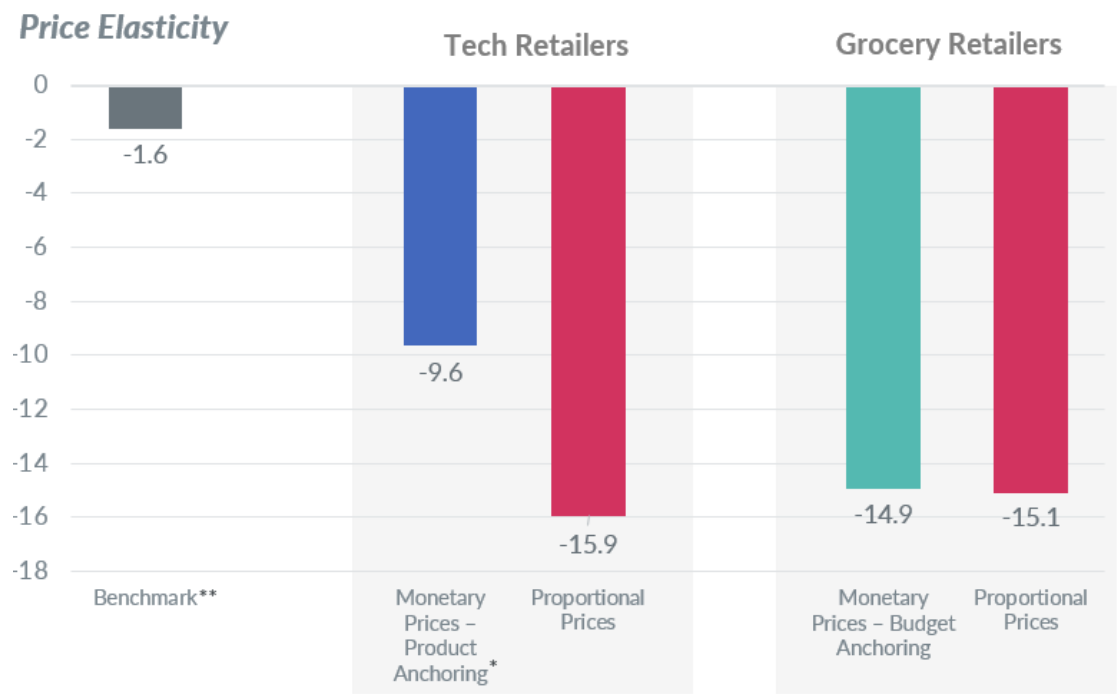


Price Elasticities

When we calculate the average price elasticity across all brands tested in the conjoint for each approach, we see that the values are very high for all approaches compared to our benchmark value, derived from various brand-price conjoint studies conducted for manufacturers’ brands across multiple sectors (Figure 11). This indicates that consumers are very sensitive to changes in retailer prices.

For grocery retailers, we see that proportional prices and monetary prices using budget anchoring give similar price elasticities. But when we look at technology retailers, monetary prices using product anchoring have a notably lower elasticity than proportional prices.

Figure 11: Average Price Elasticities of Brands Tested



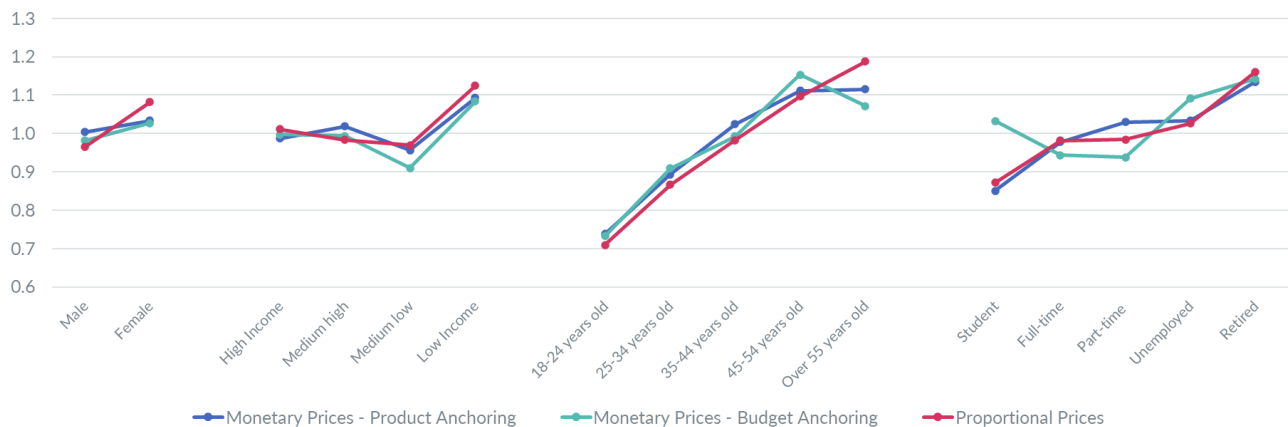
Although the overall magnitude of the price elasticities differ across the methods, we see that the price elasticities we obtain for each brand from the monetary prices and proportional prices approaches are highly correlated (Figure 12). For technology retailers, there are a couple of outliers, but generally, although the magnitudes are different, the relative price elasticities across brands are very similar for both monetary prices using product anchoring and proportional prices. For grocery retailers, the price elasticities we obtain for each brand from monetary prices using budget anchoring and proportional prices are also highly correlated, but the relationship is not as strong.

Figure 12: Correlation of Price Elasticities Between Monetary and Proportional Prices



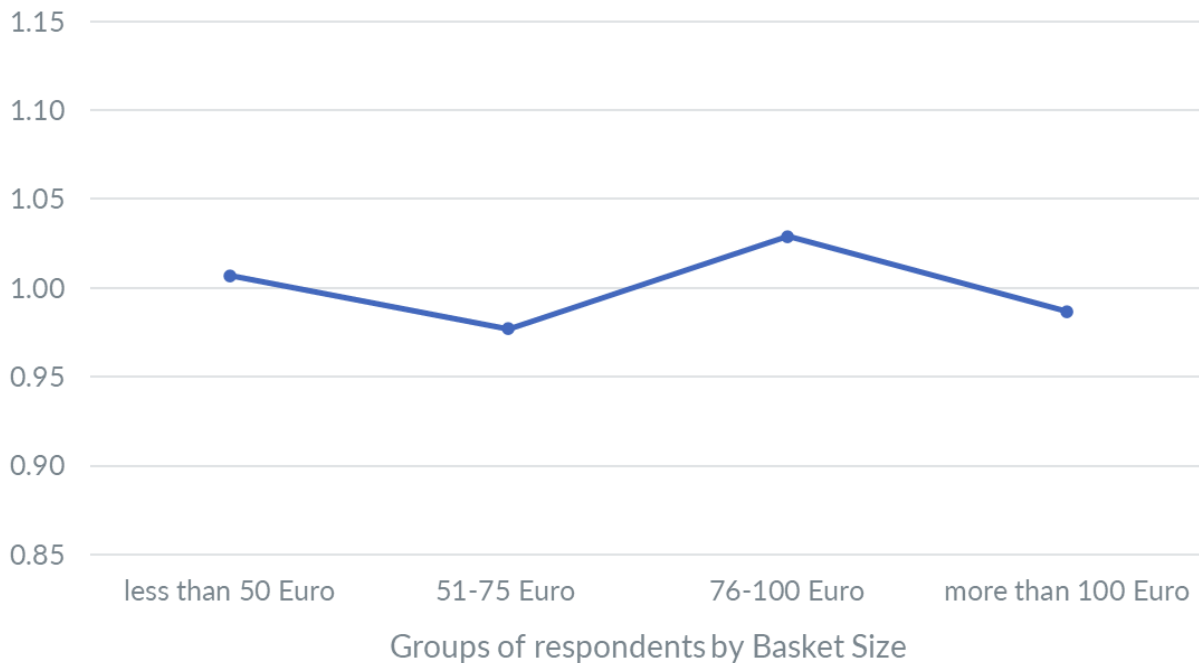
When we look at the price elasticities across different socio-demographic respondent groups, they generally align with what we would expect (Figure 13). However, the monetary prices using budget anchoring approach provides some results that are less intuitive. For example, the price elasticities are lower than expected for the over-55-years-old group and higher for students compared to the other approaches.

Figure 13: Price Elasticities by Socio Demographic Profile



In the monetary prices using budget anchoring approach, respondents state what their typical weekly spend is on grocery shopping and the prices we show to respondents are based on this typical spend. When we split the sample of respondents by their typical weekly spend or basket size and calculate the average price elasticities across brands for each respondent group, we see that the price elasticities are similar for all groups of respondents with different typical weekly spends (Figure 14).

Figure 14: Price Elasticities by Basket Size

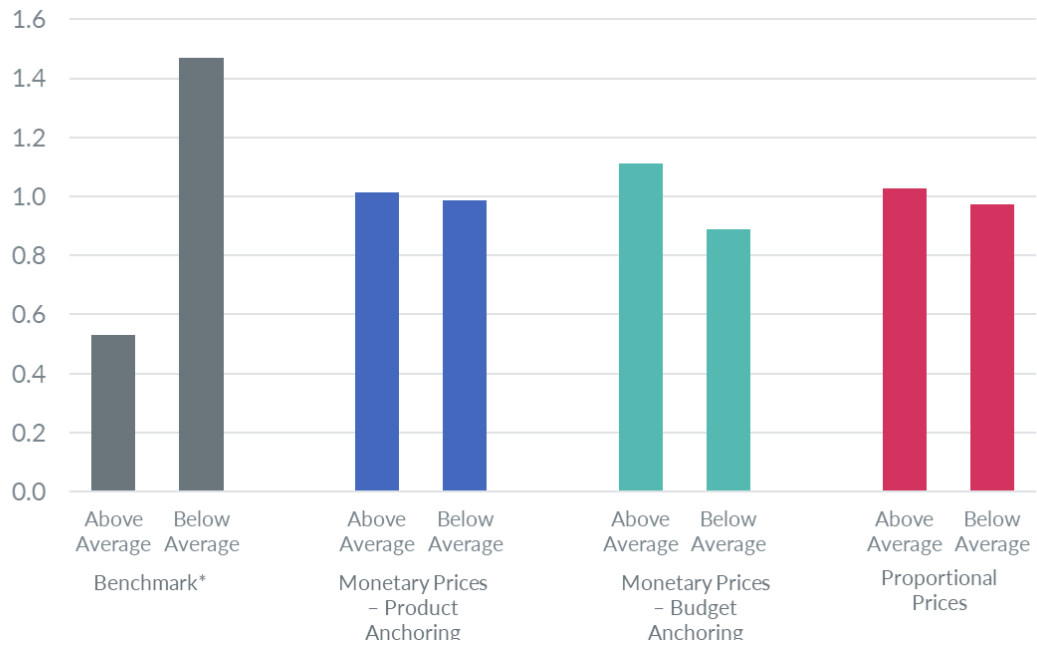


Price Elasticity values are indexed to average 1. The higher the index the higher the price elasticity.

We have seen in previous conjoint studies, run for manufacturer brands in many different categories and countries, that there is strong correlation between the price a brand charges and the brand's price elasticity. Brands that have a price higher than average have a much lower price elasticity. It is the low-price elasticity that allows them to charge a higher price because if you want to be able to increase your prices, you first need to make your customers less sensitive to price changes.

However, we do not observe the same relationship in our approaches tested in the retail industry (Figure 15). Price elasticities are not lower for retailers with a higher-than-average price. The monetary prices using budget anchoring approach shows more expensive brands to have higher elasticities, which is somewhat counter-intuitive.

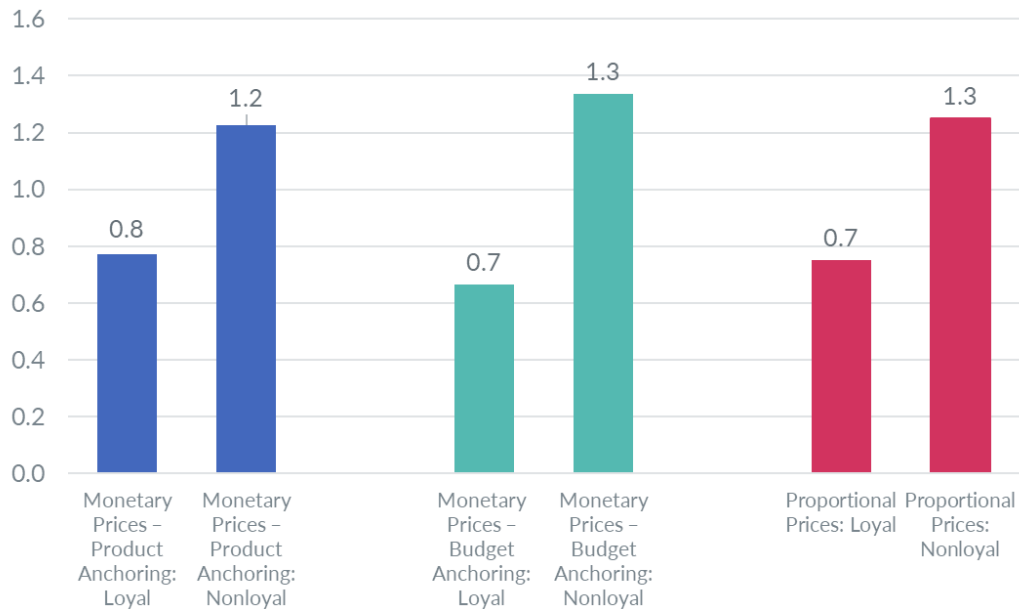
Figure 15: Price Elasticity by Retailer Price Segment



Price Elasticity values are indexed to average 1. The higher the index the higher the price elasticity.

When we look at how the price elasticities differ between loyal and non-loyal customers, as expected, we see that for all three methods, non-loyal customers have a higher price elasticity than loyal customers (Figure 16).

Figure 16: Price Elasticity by Loyal vs. Non-loyal Shoppers

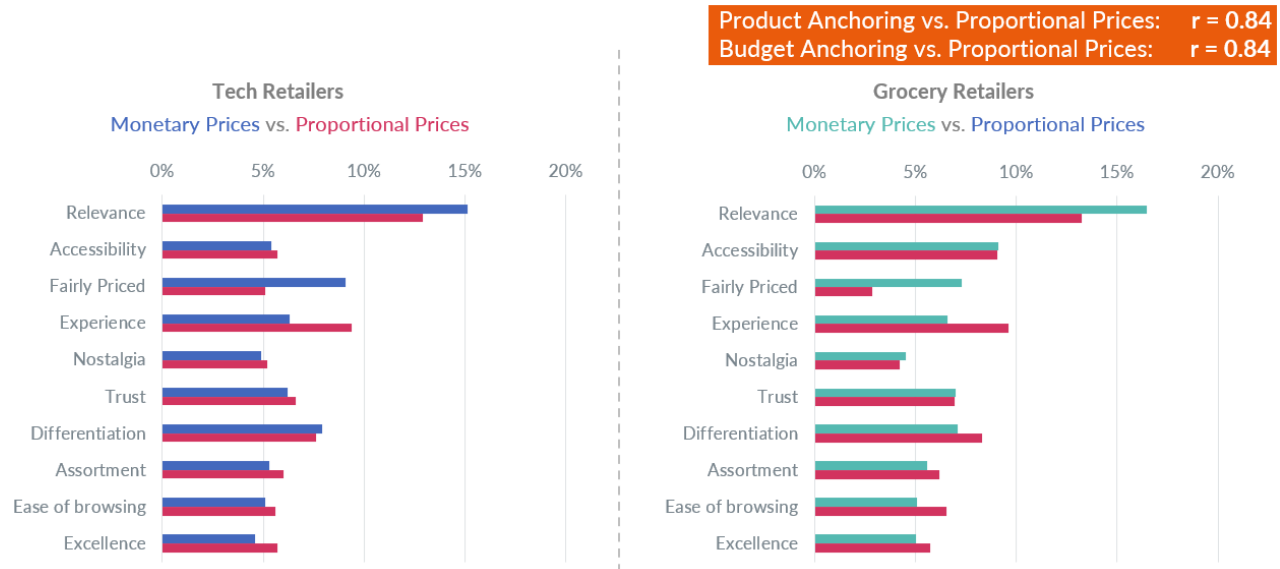


Price Elasticity values are indexed to average 1. The higher the index the higher the price elasticity.

Key Drivers and Switching Behaviour

We looked at the key drivers of preference and saw they were very similar across the three methods (Figure 17). The same is true with the drivers of loyalty, with little difference between the methods. We also looked at how shoppers switch between brands and the switching behaviour was again very similar across the methods. The switching patterns obtained from the proportional prices approach have a 0.93 correlation with the switching patterns obtained from the monetary prices using product anchoring and 0.79 correlation with the switching patterns obtained from monetary prices using budget anchoring.

Figure 17: Key Drivers of Share of Preference



DISCUSSION

Data Quality

All methods offer a very good respondent experience and have good internal validity. The fact that the monetary prices with product anchoring approach had more respondents choosing the “none” option and we identified a much higher proportion of “bad” respondents compared to the other approaches is likely because respondents were asked about a specific product category, such as an electric toothbrush. In the other approaches the category was defined more broadly. Although respondents would not reject the more-specific product category, the respondent may not be looking to make a purchase in the near future and hence they were less engaged with the conjoint exercise and/or chose the “none” option more often.

Shares of Preference vs. Real World Market Shares

All methods yield accurate estimates of retailer demand, which are closely aligned with market shares. The monetary approaches overstate the shares of premium retailers but understate the shares of low-cost retailers. This is likely because all retailers are shown at the same price,

and the respondents assume equal pricing across retailers, overlooking the real-world price differences. This over and under estimation is less apparent for technology retailers, potentially because of the lower price differentiation across the retailers.

The proportional prices approach predicts more accurately the market shares, particularly for grocery retailers, because it anchors respondents to what they know about the market and the prices they would expect at each retailer. For example, UK respondents would know Aldi and Lidl are low-cost retailers and they would expect to pay less in these stores compared with most other retailers. Similarly, they would expect to pay more in more premium retailers such as Waitrose and Marks & Spencer (M&S). When respondents make their choices in the conjoint exercise, these differences in price perceptions are taken into account, leading to shares of preference that are very close to real-world market shares.

Price Elasticity

The average price elasticities of retailers we obtain from all approaches are very high, meaning consumers are very sensitive to changes in price. If a retailer increases its prices, consumers easily switch to a competitor. The elasticities are much higher than what we normally see amongst manufacturer brands, where consumers are often willing to pay more for certain brands. But in the retailer world, shoppers seem to be unwilling to pay more for the same products at different retailers. In many cases, you are getting the exact same product, so why would you pay more?

It is also likely that all methods overestimate the price elasticities. In the conjoint exercise we are making all respondents aware of all prices of retailers whereas, in reality, shoppers may not be aware of all prices in all retailers. We also make it very easy for respondents to switch between retailers. They just have to click on a different option on the screen when in real life they would have to go to a different physical store or visit a different website and locate the product again, which requires much more effort.

We see that the size of the price elasticities for monetary prices with budget anchoring and proportional prices are exaggerated further and are notably higher than monetary prices with product anchoring. We think monetary prices with budget anchoring may amplify shoppers' sensitivity to price changes further because respondents are anchored to what they normally pay for their shopping. For example, if they know that normally they pay say €100 Euros for their weekly shopping, when you show them a price higher than \$100, the respondent feels like they are losing out. They therefore switch to another retailer.

The same anchoring effect is likely true for proportional pricing, since prices are relative to what a respondent expects to pay. When you show a respondent a price that is 10% higher than what they expect to pay, they again feel like they are losing out. In addition, respondents may perceive percentage reductions as price discounts, which are particularly attractive, and encourage a higher level of switching.

However, it could be that the lower product anchoring elasticity is, to some extent, a result from the tested price range being wider. For monetary prices with product anchoring, we tested a 20% price variation compared to a 10% price variation for the other methods and we don't know the impact this had.

Interestingly, the amount respondents typically spend on groceries doesn't influence how sensitive they are to price changes. This suggests that it is the relative price difference versus what you normally pay that matters, not the absolute magnitude of the price difference.

Ability to Answer Key Strategic Business Questions

Despite the varying magnitudes of price elasticities across methods, the patterns of price elasticities are consistent across brands and respondent groups, meaning the managerial insights you would deliver to the client would be similar irrespective of the method used. Monetary prices with budget anchoring reveals the largest differences from the other methods. However, the reasons for these discrepancies remain unclear.

All three methods can reliably identify loyal customers and provide valuable insights into how their behaviour is affected by price changes versus non-loyal customers. In addition, all methods can effectively identify the drivers of loyalty and preference, as well as how shoppers switch between retailers. Each method demonstrates consistency across various brands and aligns with our knowledge of the markets we studied. However, while the choice between methods may not significantly impact managerial recommendations, the monetary prices with budget anchoring approach again proves to be the least consistent with the other methods and exhibits some counterintuitive results.

Unfortunately, none of three methods successfully measure a brand's ability to charge a higher price than the competition. We do not see that more premium brands have a lower price elasticity, like we do amongst manufacturer brands. However, this may be due to characteristics of the market, rather than limitations of the approaches themselves. The retail markets we studied are very price sensitive and shoppers are not willing to pay more for one retailer over another. But it is clear that we cannot use price elasticities to reliably measure the pricing power of brands.

CONCLUSION

Conjoint analysis is a powerful methodology often used in tactical optimization of product prices and features. Here we have demonstrated that conjoint analysis can also be used to answer many higher-level strategic questions. Both monetary prices and proportional prices effectively assess brand preference, loyalty, switching patterns, and key drivers of preference and loyalty. However, we recommend using proportional prices to answer such questions for the following reasons. Firstly, the shares of preference from proportional price approaches are just as close, or closer, to market shares than the monetary versions. Approaches using proportional prices also answer other key business questions just as well as monetary methods, if not better.

In addition, proportional prices are much simpler to set up than monetary prices. No prices for different brands need to be sourced, which is often difficult because brands sell many different products. With proportional prices, only a simple brand list is required. Monetary prices using budget anchoring requires asking respondent about their typical spend prior to the conjoint exercise and uses that information to create the prices shown to respondents. This makes the survey more difficult to script. Monetary prices using product anchoring requires allocating respondents to a different conjoint exercise based on which products they would not reject. This

again increases the complexity as multiple conjoint exercises need to be scripted and, crucially, a much larger sample size is needed to ensure enough respondents complete each exercise. In contrast, the proportional prices conjoint exercise is very straightforward to set up.

Due to its simplicity in set up and high accuracy, proportional prices may offer a good solution to many other types of conjoint studies. The general wisdom is that it is better to be specific with features and prices you specify in conjoint exercises. But sometimes this can introduce more complexity that makes setting up and analysing the conjoint difficult. There may be cases where using a simpler approach using proportional prices would provide sufficiently accurate results whilst keeping everything far less complicated. This would help to increase speed of delivery and keep project costs down, potentially increasing the cases we can offer conjoint analysis to clients. However, more validation is needed to assess how well proportional prices performs in different circumstances, but we hope this paper inspires you to consider using proportional prices in your future conjoint analysis projects.



Alexandra Chirilov



James Pitcher

DESIGN AND MODELING CONSIDERATIONS WITH A DOMINATING ATTRIBUTE

JOE JONES
LISA MARIN
ADELPHI RESEARCH

1.0 BACKGROUND

Choice-based conjoint analysis exercises (CBC) are commonly used in marketing fields by researchers to understand decision-making processes. The factors that drive the decision process will vary greatly from industry to industry. However, industries often utilize branding and price as a part of the CBC. These are critical components in understanding the choice process but can sometimes pose problems during the analysis stage. Specifically, these problems occur from being dominating factors compared to other features in the CBC.

Within a CBC context, a dominating attribute is a key factor in the decision process and will likely overshadow other factors. For instance, a respondent may look at a series of concepts and choose the cheapest option, ignoring all other aspects. While this is an important part of the decision process, little to no information is gained beyond price. In certain cases, it is possible to remove the dominating attribute to understand the decision process within other features, but this is not always possible. Specifically, if revenue, profit, and forecasting metrics are desired, price should not be dropped from the CBC.

One industry where excluding a dominating attribute is unlikely is healthcare and pharmaceuticals. For instance, price, safety, efficacy, time to results, etc., are pertinent to include in pharmaceutical product designs. In the presence of a dominating attribute, special considerations for the design and modeling are required to better accommodate the non-dominating attributes.

Specifically, researchers must find a way to get respondents to look beyond the dominating attribute. A simple way to do this is by making multiple concepts show the same level of the dominating attribute within a task. Continuing with the price example where a respondent chose the cheapest option, if the respondent were to see two profiles with identical prices, they would be forced to look beyond price and examine the other features in the CBC. When designing CBCs, how often the same level is shown together is called overlap.

The present study examined how altering the amount of overlap in the CBC design impacts the decision process. We hypothesized that having more overlap will reduce the importance of a dominating attribute, leading to increased accuracy and understanding of the non-dominating trade-offs.

2.0 METHODS AND MATERIALS

2.1 Attribute & Levels Grid

The objective of a CBC is to elicit preferences or values over a range of attributes and levels that define various treatment profiles in the choice tasks that participants are asked to complete. The identification, selection and testing of attributes and associated levels of these attributes, as

summarized in an Attributes & Levels (A&L) grid was the first step in the conduct of this study (Mesana et al., 2023).

The development of the A&L grid began with the identification of elements involved in patient decision-making in the purchase of over-the-counter pain medication. The intention was to include a dominating attribute that would be of universally paramount importance to respondents. Price was selected as the dominating attribute with levels spanning a broad range (\$5.00–\$100.00 USD) to ensure high elasticity and overshadow the other attributes in the grid. Additionally, administration was selected as a primary attribute upon which conditional price attributes were built with the theory that different administration methods could yield different willingness-to-pay (Rahimi, Rasekh, Abbasian, and Peiravian, 2018).

It was decided to limit levels to three or four per attribute, in accordance with best practice (Bridges et al., 2011) and to ensure a manageable number of CBC tasks without compromising the standard errors of the coefficient estimates (<0.05), given the available sample (see section 2.3). It was ensured that levels were mutually exclusive to facilitate differentiation between profiles. To minimize potential ambiguity for participants, the use of ranges to define individual attribute levels was avoided (Mesana et al., 2023).

2.2 Quantitative Choice-Based Conjoint (CBC) Survey

Applied to healthcare decision-making, CBCs assume that a treatment is decomposed into its constituent parts: attributes (e.g., route of administration, side effects, safety warnings), levels per attribute (e.g., patch and oral capsule, needs to be refrigerated and room temperature), and that the utility of a treatment to an individual patient is a function of the attribute levels that define the treatment (Ryan and Farrar, 2000). The objective of CBC was to quantify the trade-offs individuals make between attributes when making choices between two hypothetical treatment options (Mesana et al., 2023). This included determining the relative importance of each attribute among attributes tested (Sawtooth Software Inc., 2013). By independently varying the attribute and levels of the treatment choice tasks and observing the pattern of responses, the impact of a change in each attribute on the probability of choosing an alternative can be inferred statistically (US Food and Drug Administration, 2016; Overbeeke, Vanbinst, Jimenez-Moreno, and Huys, 2020).

The CBC required participants to choose between three alternatives determined by an experimental design, as well as an opt-out alternative which was specified as “None of these.” The CBC was part of a larger survey instrument that was hosted online. The online survey included five components:

- Eligibility Screener
- Baseline characteristic questions about current treatment options
- Discrete Choice Experiment
- Experience with cannabis treatment alternatives
- Demographics

The final A&L grid (Figure 1) consisted of 8 attributes and a maximum of 4 levels for each attribute designed in accordance with best practice for CBC studies. Recognizing the potential influences of efficacy and dosing on treatment preferences, these elements were controlled for in the CBC by fixing the efficacy and dosing text to be constant across treatment profiles. Efficacy for each profile was listed as “3 times more effective than over-the-counter ibuprofen” while dosing was listed as “5 mg per dose.”

Eight versions of the CBC design were created using Sawtooth Software's Lighthouse Studio 9.14.2. Four main effects designs were generated using each of the random task generation methods available in the software: complete enumeration, shortcut, random and balanced overlap. The A&L grid was then modified to include 3 pricing attributes conditional on each of the modes of administration: capsule, patch, and ointment. The resulting conditional, or alternative specific design, therefore had attributes for capsule price, patch price and ointment price in place of the one price attribute in the main-effects design. Four alternative specific designs were generated using each of the random task generation methods available in Lighthouse Studio 9.14.2: complete enumeration, shortcut, random and balanced overlap.

The designs were tested by considering the efficiency of the main and alternative specific effects—that is, for each attribute, whether there are differences in coefficient (i.e., part-worth utility) estimates across levels. This provided a good approximation of the efficiency of the CBC design (Mesana et al., 2023). For each attribute and level, an approximation was made of the relative standard error of each main effect. The designs were tested using aggregate multinomial logit (MNL) using only the information about the design of the choice tasks, rather than the respondent's answers to calculate the relative standard error estimates (Guest, Bunce, and Johnson, 2006). The pattern of their relative magnitudes with respect to one another, rather than a precise estimate of each standard error, is important for testing the design (Sawtooth Software, 2008). These standard error estimates represented the accuracy of the estimates given the design parameters available, with a lower value indicating that there is sufficient design efficiency (orthogonality and balance) within the design (Mesana et al., 2023). These estimates use simulated responses to the trade-off tasks, meaning that the standard error is only an approximation to test the design. Lighthouse Studio 9.14.2 (Chrzan and Orme, 2000) recommends that these priori estimates of standard errors for main effects are <0.05 and for alternative-specific effects <0.1 , as this allows for a robust design to test differences across levels for each attribute (Chrzan and Orme, 2000, Sawtooth Software, 2008). Each design consisted of 10 random tasks and 2 randomly generated holdout tasks, also known as fixed tasks, that were consistent across each of the 30 versions of every design and across the 8 designs. The number of random tasks was selected to minimize respondent burden while ensuring sufficient design efficiency.

Figure 1

Attribute #	Attribute Label	Level 1	Level 2	Level 3	Level4
1	Brand Type	Generic	Name Brand		
2	Safety Warnings	Do not take with alcohol	Can cause drowsiness (no restrictions on activity)	None	
3	Side Effects	10% of patients experience mild diarrhea	10% of patients experience mild rash	10% of patients experience mild headache	
4	Administration	Patch worn for at least 4 hours	Topical ointment (dries in 5 minutes)	1 capsule taken orally	1 capsule taken orally with food
5	Recommended Administration Frequency	Twice per day	Once per day		
6	Storage	Needs to be refrigerated	Room temperature		
7	Cost for 30-Day Supply (USD)	\$100	\$50	\$20	\$5
8	Availability	Retail only with restricted access (staff assistance needed to unlock medication)	Online and retail (no restrictions to access)		

2.3 Sample

Participants (N = 1,817) were recruited from Dynata, LLC. The sample comprised 37% males and 63% females, aged between 18 and 89 years (M = 53.01, SD = 16.63). All participants gave informed consent before participating in the study. Participants were screened to ensure they had experienced headaches, body aches or pains in the past 12 months but have not been prescribed prescription medication to treat their pain. Participants were assigned to 1 of 8 patient types based on their response to their assigned sex at birth (male or female) and pain frequency. Pain frequency was bucketed into 4 categories: rarely, occasionally, often, and regularly. Therefore, the 8 patient types were as follows:

- Male Rarely
- Male Occasionally
- Male Often
- Male Regularly
- Female Rarely
- Female Occasionally
- Female Often
- Female Regularly

Within each patient type, CBC designs were assigned based on least filled cell. Within each design, one of 30 versions was assigned based on least filled cell. A minimum sample of 225 US-based participants completed each of the 8 designs. The minimum sample of 225 participants was determined to be a sufficient sample size to allow a manageable number of CBC tasks without compromising the statistical precision of the resulting estimates. The number of respondents in each design is outlined below in Figure 2.

Figure 2: Participants Per Design

	Main Effects	Alternative-Specific
Complete Enumeration	N=227	N=229
Shortcut	N=226	N=226
Random	N=225	N=230
Balanced Overlap	N=226	N=228

2.4 Analysis

2.4.1 Data Checking

The conjoint exercise started with a dummy task and an open-ended follow up question which asks the respondent for their rationale as to why they have picked the particular response. The dummy task was set up so that one treatment contains fewer desirable levels across all attributes. If respondents select this concept, they were asked to supply their rationale, and this was be flagged to Adelphi for closer review. During fieldwork quality control was ensured through Adelphi's Advanced QualityChecker (Kennedy, Marin, Hallworth, Mellor, and Hughes, 2022). QualityChecker employs three key measures; these included checking speed of responses, flatlining (always choosing the none option for example), and patterned responding, i.e., alternating responses left then right. Respondents highlighted by these checks were scrutinized and replaced as necessary.

2.4.2 Utility Estimation

Hierarchical Bayesian (HB) estimation enables us to obtain individual level part-worth utility estimates unlike pooled logit-based methods which provide aggregate estimates (Orme, 2000; Rao, 2014). The Hierarchical Bayes model has two levels. At the higher level, we assume that individuals' parameters are described by a multivariate normal distribution (Rao, 2014). At the lower we assume that, given an individual's part-worth utilities, their probability of choosing a particular profile is governed by a multinomial logit model (Orme, 2000).

Aggregate estimation models confound heterogeneity and noise. By modeling individuals rather than the average, HB can separate heterogeneity from noise. This leads to more stable, accurate models whether viewed in terms of individual or aggregate level performance (Orme, 2000).

Hierarchical Bayesian estimation is applied to CBC data to estimate the relative value each respondent puts on an attribute level. The utility function used is the part-worth model (Allenby, Arora, and Ginter, 1995; Hauber et al., 2016; Sawtooth Software, 2019; Soekhai, de Bekker-Grob, Ellis, and Vass, 2019). This model represents attribute-level utilities by a piecewise linear curve. The part-worth model reflects a utility function that defines a different utility value for each of the levels of an attribute (Mesana et al., 2023). Attribute levels are modeled as categorical variables and effects coded. The part-worth is estimated using effects coding are generally easier to interpret than for dummy coding.

The values from the HB estimation are called part-worth utilities (i.e., attribute-level utilities). A low utility indicates less preference; a high utility or desirability indicates more preference. A negative utility does not necessarily mean that a respondent did not like a given level just that relative to other levels within the same attribute, it was less preferred to other levels (Mesana et al., 2023). You cannot directly compare the part-worth utility of a level from one attribute to the part-worth utility of a level from another attribute, but you can directly compare part-worth utilities for levels within the same attribute (Mesana et al., 2023). Raw part-worth utilities from the HB estimation are zero-centered (i.e., the utilities sum to zero within each attribute).

2.4.3 Restructuring the Designs for Analysis

Each CBC cell underwent two sets of analyses. The first analysis was conducted on each design, as it was generated by Lighthouse Studio 9.14.2. The second analysis was conducted on a design that had been restructured in R Studio version 4.1.3.

Each main effects design was transformed into an alternative-specific model by recoding the design from 1 unconditional price attribute to 3 price attributes conditional on administration (patch price, ointment price, capsule price) to run the estimation. A hypothetical example of this transformation is found in Figure 3 below.

Figure 3:
Main-Effects from Lighthouse Studio 9.14.2

Version	Task	Concept	Administration	Cost
1	1	1	1	2
1	1	2	2	4
1	1	3	4	3

Restructured Design for Alternative-Specific Model

Version	Task	Concept	Administration	Cost_Admin1	Cost_Admin2	Cost_Admin3/4
1	1	1	1	2	0	0
1	1	2	2	0	4	0
1	1	3	4	0	0	3

Conversely, each alternative-specific design was transformed into a main effects model by recoding the design. The 3 price attributes conditional on administration (patch price, ointment price, capsule price) were collapsed down into a single unconditional price attribute to run the estimation. A hypothetical example of this transformation is found in Figure 4 below.

Figure 4:
Alternative-Specific Design from Lighthouse Studio 9.14.2

Version	Task	Concept	Administration	Cost_Admin1	Cost_Admin2	Cost_Admin3/4
1	1	1	3	0	0	3
1	1	2	4	0	0	4
1	1	3	2	0	3	0

Restructured Design for Main Effects Model

Version	Task	Concept	Administration	Cost
1	1	1	3	3
1	1	2	4	4
1	1	3	2	3

Note, the alternative-specific design in the figure above (Figure 4) utilizes an analogous approach to L^{MN} design (Chrzan and Orme, 2000). Specifically, we are building unique cost attributes and the algorithm does not necessarily know that level 3 for “Cost_Admin2” is the same as level 3 of “Cost_Admin3/4.” Thus, there is some additional overlap artificially built into the design, which can be recognized when restructured into a main-effects design (i.e., second table of Figure 4).

2.5 Evaluation Criteria

2.5.1 Root Likelihood (RLH)

Root likelihood, or RLH, was used to measure the goodness of fit for each model. RLH is computed by simply taking the n^{th} root of the likelihood, where n is the total number of choices made by all respondents in all tasks. Therefore, RLH is the geometric mean of the predicted probabilities. If there were k alternatives in each choice task and assuming no information about part-worth utilities, we would predict that each alternative would be chosen with probability $1/k$, and the corresponding RLH would also be $1/k$. If the fit were perfect, RLH would be 1 (Sawtooth Software, 2013).

2.5.2 Holdout Task Hit Rate

The accuracy of each model was assessed by evaluating the validity of the data. Two randomly generated holdout tasks were included in the CBC that were not used in the HB estimation. These tasks provided an indication of validity, measured by the utility estimates’ ability to predict choices not used in their estimation (Orme and Johnson, 2015). Due to sample limitations, only within-sample holdout task choices were assessed.

2.5.3 Attribute Importance

The relative importance of the individual attributes for each model were calculated two different ways. The first method of calculation used the part-worth utilities to calculate relative importance using a method based on the variability explained by the various attributes in the CBC (Hauber et al., 2016). The aim of this is to calculate how much difference each attribute could make in the total utility of a disease state. Using the individual-level part-worth utilities, the range (maximum minus minimum) was taken for each attribute for each participant. These ranges were then reproportioned (to a percentage) across the attributes to sum to 100%, giving the relative importance of each attribute. The relative importance for each respondent is calculated and then averaged to arrive at relative importance for each attribute.

The second method of calculating relative attribute importance utilized share of preference simulations. For these simulations, pre-defined profiles (defined in terms of selected levels on each of the attributes included in the study design) were created and organized into scenarios. The simulation method used preference share, which was calculated by summing the part-worth utilities of each attribute level corresponding to the grid. The sum of the part-worth utilities for each profile was then subjected to the exponential transformation, before rescaling the resulting numbers so that they sum to 100%. The preference share for each profile indicates the probability of choosing a device relative to other alternatives in a given scenario (Chrzan and Orme, 2000, Sawtooth Software, 2008).

To calculate attribute importance, a series of preference share simulations were run comparing the test profile to the base-case profile. The levels of each attribute of the test profile were varied one at a time while holding all other attributes constant to find the range (maximum minus minimum) of shares for each attribute individually. In the alternative-specific models, the primary and conditional attributes, administration and price, respectively, were considered as a single with all other attributes held constant. The ranges were then rescaled to sum to 100% giving the relative importance for each attribute.

Since the A&L grid was hypothetical and not intended to resemble any real-life market scenario, base-case profiles were defined as the attribute levels used in each of the 6 holdout task concepts (as the designs included 2 holdout tasks with 3 concepts per task). Therefore, 6 sets of preference share simulations and resulting attribute importance scores were calculated for each model. The reported simulated based importance scores for each attribute for each model were the mean of the attribute importance scores that resulted from each set of simulations.

3.0 RESULTS

3.1 Model Fit

We first compared RLH across the 16 models and found alternative-specific models yielded the highest fit metrics. This effect was seen regardless of the design used, where a main effect design restructured and analyzed as an alternative specific model showed comparable scores to an alternative specific design and model. On the other hand, main effect models using an alternative specific design showed the worst RLH value.

With respect to design algorithms, balanced overlap designs resulted in the best model fit, with complete enumeration and random designs having the worst model fit. These results indicate using design algorithms where more overlap is used leads to better RLH values. See Table 1 below for all RLH values.

Table 1

Design	Model	Complete Enumeration	Shortcut	Random	Balanced Overlap	Overall Model
Main Effects	Main Effects	67.3%	68.7%	65.4%	69.3%	67.6%
	Alternative Specific	70.0%	70.8%	67.7%	72.1%	70.1% ^{A-ME}
Alternative Specific	Main Effects	63.6%	67.3%	62.8%	70.3%	66.0%
	Alternative Specific	67.8%	70.9%	66.4%	73.3%	69.6% ^{A-ME}
Overall Task Generation		67.2%	69.4%	65.6%	71.2% ^{CE, Ra}	

Note: Superscript indicates where values are significantly different from other cells at the 95% confidence interval.

3.2 Accuracy

Next, we examined accuracy across the 16 models and found main effects or alternative specific designs and modeling had no impact on accuracy. However, the design algorithm did show a significant effect on accuracy. Specifically, the design algorithms that incorporate the most overlap, random and balanced overlap, yielded the most accurate models. This is in line with the hypothesis that having more overlap leads respondents to looking beyond the dominating attribute, thus leading to a better understanding of non-dominating trade-offs, and increased accuracy. It should also be noted that complete enumeration used with alternative specific designs yielded the worst accuracy. See Table 2 below for all accuracy rates.

Table 2

Design	Model	Complete Enumeration	Shortcut	Random	Balanced Overlap	Overall Model
Main Effects	Main Effects	74.7%	70.8%	75.3%	72.8%	73.4%
	Alternative Specific	72.9%	70.8%	74.7%	73.5%	73.0%
Alternative Specific	Main Effects	66.8%	72.6%	72.8%	73.7%	71.5%
	Alternative Specific	66.2%	72.8%	71.7%	73.7%	71.1%
Overall Task Generation		70.1%	71.7%	73.6% ^{CE, Sh}	73.4% ^{CE, Sh}	

Note: Superscript indicates where values are significantly different from other cells at the 95% confidence interval.

3.3 Importance

As stated before, importance scores were examined two different ways for the main effects models. Specifically, individual importance scores were calculated via utilities and simulations where change in share of preference was assessed. For alternative specific models, given the three different pricing attributes, only the simulation importance scores were calculated. Results showed an interesting pattern when comparing importance scores from utilities versus simulations. Particularly, importance scores for the dominating attribute calculated from the utilities were much lower than the simulations. This was true in all cases regardless of design, and the difference was more prominent in algorithms with less overlap.

When comparing importance scores on the dominating attribute between main effects and alternative specific designs, we see alternative specific designs showing the lowest importance scores regardless of how the design was analyzed. In addition, we also see the algorithms with the most overlap providing the lowest dominating importance scores. Like accuracy metrics, the increased overlap is forcing respondents to look beyond the dominating overlap in the decision process, thus leading to increased importance scores in the non-dominating attribute. However, we would like to note that there was a large amount of variability in the importance scores and none of the results were significantly different, but the trend does suggest increased overlap influences importance scores. See Table 3 for the importance scores.

Table 3

Design	Model	Complete Enumeration	Shortcut	Random	Balanced Overlap	Overall Model
Main Effects	Main Effects (utility)	54.8%	49.9%	53.2%	52.8%	52.7%
	Main Effects (simulation)	71.4%	62.6%	69.6%	57.9%	65.4%
	Alternative Specific (simulation)	70.7%	56.7%	69.0%	53.4%	62.5%
Alternative Specific	Main Effects (utility)	49.8%	51.1%	50.8%	54.5%	51.6%
	Main Effects (simulation)	59.6%	58.5%	62.5%	61.6%	60.6%
	Alternative Specific (simulation)	54.0%	55.3%	57.3%	59.8%	56.6%
Overall Task Generation		63.9%	58.3%	64.6%	58.2%	

4.0 SUMMARY/RECOMMENDATION

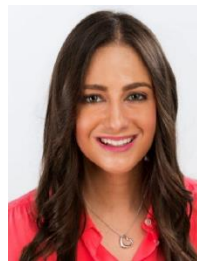
We recommend using alternative-specific designs created with the balanced overlap random task generation method. Designs created using balanced overlap have the highest model fit, in terms of RLH. Meanwhile, alternative specific designs yield higher non-dominating importance scores when using share of preference simulations to calculate importance. These alternative

specific designs created with balanced overlap random task generation boost overlap in the levels shown of the dominating attributes, which help increase the non-dominating attribute importance scores by forcing the model to focus on non-dominating attributes.

At the modeling stage, we recommend exploiting both main effects and alternative specific models and leveraging insights from both. While alternative specific models yield higher RLH than main effects models, it's important to think practically about how results are going to be employed to determine where and to what degree overlap should be built into the design. At this juncture, we're not able to speak to the true importance of the dominating attribute and, therefore, cannot say which designs and models inflated or deflated dominating attribute importance. However, plans for future research include running simulated data based on current findings to assess patterns and arrive at the true importance.



Joe Jones



Lisa Marin

REFERENCES

- Allenby, G. M., Arora, N., and Ginter, J. L. (1995). Incorporating prior knowledge into the analysis of conjoint studies. *Journal of Marketing Research*, 32(2), 152–162.
- Bridges, J. F., Hauber, A. B., Marshall, D., Lloyd, A., Prosser, L. A., Regier, D. A., . . . and Mauskopf, J. (2011). Conjoint analysis applications in health—a checklist: a report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. *Value in health*, 14(4), 403–413.
- Chrzan, K., and Orme, B. (2000). An overview and comparison of design strategies for choice-based conjoint analysis. *Sawtooth software research paper series*, 98382, 161–178.
- Guest, G., Bunce, A., and Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field methods*, 18(1), 59–82.
- Hauber, A. B., González, J. M., Groothuis-Oudshoorn, C. G., Prior, T., Marshall, D. A., Cunningham, C., . . . and Bridges, J. F. (2016). Statistical methods for the analysis of discrete choice experiments: a report of the ISPOR conjoint analysis good research practices task force. *Value in health*, 19(4), 300–315.
- Kennedy, C., Marin, L., Hallworth, P., Mellor, P., and Hughes, O. (2022). *Patient preferences for targeted therapies in metastatic melanoma: Statistical analysis plan* [Manuscript submitted for publication].
- Mesana, L., Chen, K., Mason, B., Clifford, M., Randhawa, S., Gater, A., . . . and Jones, J.W. (2023). *Patient Preferences for Targeted Therapies in Metastatic Melanoma* [Manuscript in preparation].

- Orme, B. (2000). Hierarchical Bayes: why all the attention. *Quirk's Marketing Research Review*, 14(3), 16–63.
- Orme, B., and Johnson, R. (2015). Including holdout choice tasks in conjoint studies. *Sawtooth Software: Research Paper Series*.
- Rahimi, F., Rasekh, H. R., Abbasian, E., and Peiravian, F. (2018). A new approach to pharmaceutical pricing based on patients' willingness to pay. *Tropical Medicine & International Health*, 23(12), 1326–1331.
- Rao, V. R. (2014). *Applied conjoint analysis*. Springer Science & Business Media.
- Ryan, M., and Farrar, S. (2000). Using conjoint analysis to elicit preferences for health care. *Bmj*, 320(7248), 1530–1533.
- Sawtooth Software (2013). Sawtooth Software Technical Paper Series: The CBC System for Choice-Based Conjoint Analysis. Version 8. *Sawtooth Software Inc.*
- Sawtooth Software. (2008). The CBC advanced design module technical paper sawtooth software technical paper series 1–31.
- Sawtooth Software. (2013). The CBC/HB System for Hierarchical Bayes Analysis and Advanced Simulation [Software manual]. Sawtooth Software.
- Sawtooth Software. (2019). Lighthouse studio manual.
- Soekhai, V., de Bekker-Grob, E. W., Ellis, A. R., and Vass, C. M. (2019). Discrete choice experiments in health economics: past, present and future. *Pharmacoeconomics*, 37, 201–226.
- US Food and Drug Administration. (2016). Patient Preference Information—Voluntary Submission, Review in Premarket Approval Applications, Humanitarian Device Exemption Applications, and De Novo Requests, and Inclusion in Decision Summaries and Device Labeling. Guidance for Industry. *Food and Drug Administration staff, and other stakeholders*, 2017.
- Van Overbeeke, E., Vanbinst, I., Jimenez-Moreno, A. C., and Huys, I. (2020). Patient centrality in patient preference studies: the patient perspective. *Frontiers in medicine*, 7, 93.

COMPARING SYSTEM 1 PRIMING VS. MAXDIFF: WHICH APPROACH MEASURES BRAND PERCEPTIONS MORE ACCURATELY?

MICHAEL PATTERSON

SONIA HUNDAL

RADIUS GLOBAL MARKET RESEARCH

ABSTRACT

This research was designed to investigate three typical approaches to measuring subconscious, System 1 processing; namely Implicit Priming Test (IPT), paired comparison MaxDiff approach which we term the Emotional Valence Test (EVT), and an approach that combines both, the Adaptive EVT. We examined the relationship between System 1 and System 2 measures for both low emotional valence brands (Visa and MasterCard) along with high emotional valence brands (Fox News and MSNBC). We also assessed the test-retest reliability of the System 1 approaches. Our results show that while all three techniques perform well, the IPT approach warrants consideration due to its simplicity and effectiveness.

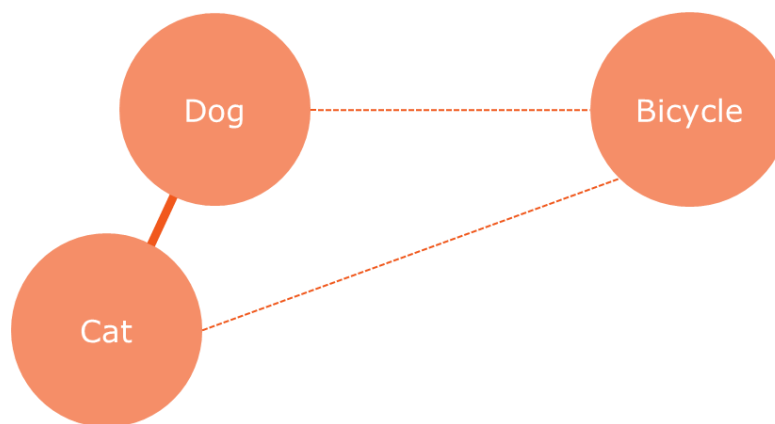
INTRODUCTION

Quantitative market research has traditionally relied on analytical, rational models for interpreting consumer behavior. However, this approach predominantly assumes System 2 processing which can be characterized as being deliberate, analytical, and relatively slow in nature. In recent years, market researchers have increasingly recognized the importance of System 1 processing—the fast, intuitive, and often subconscious part of decision-making which accounts for the vast majority of individuals' decisions and many consumers' purchase decisions (Kahneman, 2011).

The concept of System 1 processing, popularized by Nobel laureate Kahneman (2011), challenges the traditional market research model, which primarily focuses on stated preferences, logical reasoning, and conscious awareness. Unlike System 2, System 1 acknowledges that consumers are not always rational actors but are often guided by their immediate, instinctual reactions. With the realization that consumers frequently make decisions based on emotions, instincts, and subconscious biases, the assessment of System 1 processing has emerged as a compelling area of investigation within market research. An example of System 1 processing is commuting to work, and once you've arrived, not really remembering the specific route you took—it just happened automatically. Contrast that with System 2, which is a much more deliberate, cognitively involved thought process where you really trade off and think about different options before making a decision or choice. For instance, purchasing a vehicle involves a lot more consideration before making a final decision.

The inclusion of System 1 processing in market research does not discredit the value of System 2 but highlights the need for a more balanced perspective. This fusion of subconscious and analytical cognitive processing provides a more nuanced lens to interpret and predict consumer behavior.

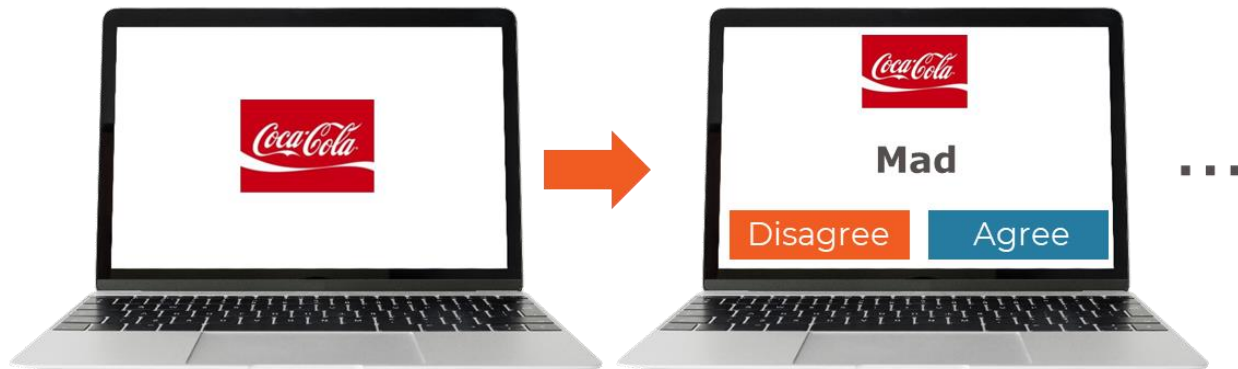
In this research we investigate two common approaches to measuring System 1 processing, and their relationship with System 2 processing. A third approach, Implicit Association Test (IAT), is not considered since we've previously found it lacking both in terms of its relationship with System 2 measures, as well as its test-retest reliability (Patterson and Frazier, 2017). Priming is a phenomenon that often occurs in the real world whereby individual's thinking or memory is influenced by information they encounter, or via positive or negative events that occur. For example, after seeing the word "cat," individuals will be faster to recognize, or readily think about, the word dog than an unrelated word such as bicycle due to stronger mental connections between cat/dog in comparison to cat/bicycle.



In addition to occurring naturally, priming can also be induced artificially using different approaches. The most common form of implicit priming in market research is the *Implicit Priming Test (IPT)* which involves exposing respondents to a stimulus (e.g., a specific brand name or advertisement) and then measuring their agreement/disagreement along with their reaction time when presented with various words consisting of different descriptors or emotional reactions. By comparing the reaction time with the combination of objects and their attributes, we are able to deduce which characteristics (e.g., beautiful, stylish, desirable, expensive, useless) or emotions (e.g., cheerful, appalled, frustrated) are more closely associated with different objects. By using the Implicit Priming Test, researchers can thus presumably uncover the "real," subconscious attitudes of respondents.

Below is an example of the Implicit Priming Test. In this case, respondents are exposed to a brand name (e.g., Coca-Cola) as the prime. They are then shown various emotions (e.g., "mad") and instructed to indicate whether they agree or disagree that they associate the emotion with the brand. Because we're trying to tap into System 1 processing, we're going to ask them to respond as quickly as they possibly can, and we're going to measure the amount of time it takes them to respond (i.e., record their reaction time). Presumably, the closer the connection of the emotion with the brand, the faster their response will be.

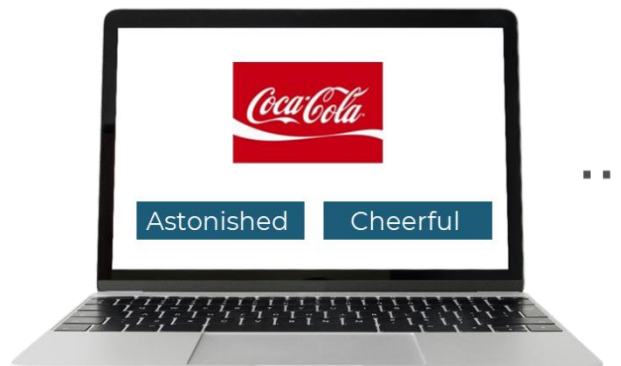
Example of the Implicit Priming Test (IPT)



Paired-comparison MaxDiff is another approach that can be used to understand how individuals perceive various stimuli. In this case, respondents can be shown a variety of different characteristics or emotional descriptors and asked which one they most and least associate with a specific object. In the case of a traditional MaxDiff exercise, respondents will most often engage in System 2 thinking since they can take the time to carefully consider and deliberate on their responses. However, if we greatly limit the amount of time that respondents have to respond along with the number of descriptors shown at a time, we can more closely approximate System 1 thinking by forcing a faster, more automatic (i.e., less cognitive) decision. We refer to this very time-constrained, paired-comparison MaxDiff approach as *Emotional Valence Testing (EVT)*.

Below is an example of an Emotional Valence Test. In this exercise, respondents are shown two emotions at a time (e.g., astonished and cheerful) and asked to indicate which one of the emotions they most closely associated with the brand (e.g., Coca-Cola). In order to induce System 1 processing, they are only given two seconds to respond. If they fail to respond within that timeframe, the screen is shown to them again later.

Example of the Emotional Valence Testing (EVT)



The use of MaxDiff and similar approaches have been previously investigated by various researchers. For example, Lipovetsky (2020) demonstrated that MaxDiff can be used to understand System 1 processing. In addition, some market research firms such as The Rational Heart measure emotional reactions by presenting pairs of words based on Plutchik's Wheel of

Emotions. When presented with the words, respondents must quickly select (within two seconds) the emotional descriptor that best represents how they feel about the test object. Given the number of emotional descriptors to be tested (24 in the case of Plutchik), respondents must go through a large number of paired comparisons in order to have stable individual-level utilities.

In our research practice, we've used both Implicit Priming Tests and Emotional Valence Tests a number of times for various brands across different categories and types of studies and have always found that both approaches work well. Notably, we find that they provide good discrimination across both brands and attributes. In the table below, we show the illustrative results from a previous study to demonstrate the differentiation we typically find.

Illustrative Results for IPT

	Brand 1	Brand 2	Brand 3	Brand 4	Brand 5	Brand 6	Brand 7	Brand 8	Brand 9	Brand 10
Att 1	81%	77%	79%	82%	77%	81%	78%	78%	77%	76%
Att 2	72%	68%	66%	71%	66%	77%	67%	71%	71%	72%
Att 3	34%	28%	35%	33%	37%	48%	34%	38%	42%	49%
Att 4	59%	59%	65%	60%	60%	78%	63%	63%	64%	64%
Att 5	83%	78%	81%	79%	78%	81%	77%	84%	77%	73%
Att 6	87%	84%	83%	81%	79%	82%	80%	83%	79%	81%
Att 7	66%	67%	70%	69%	68%	73%	62%	67%	67%	69%
Att 8	79%	74%	76%	74%	77%	81%	74%	78%	74%	77%
Att 9	86%	77%	86%	80%	78%	82%	77%	80%	78%	78%
Att 10	62%	59%	61%	59%	62%	78%	57%	62%	70%	68%
Att 11	63%	58%	66%	58%	62%	76%	55%	61%	69%	67%
Att 12	49%	46%	54%	48%	49%	68%	47%	51%	57%	64%
Att 13	56%	51%	57%	56%	54%	69%	54%	59%	64%	61%
Att 14	76%	73%	75%	71%	72%	74%	71%	75%	73%	68%
Att 15	66%	60%	66%	62%	63%	68%	62%	69%	69%	64%
Att 16	75%	65%	72%	70%	69%	67%	66%	71%	69%	73%
Att 17	73%	68%	75%	71%	66%	77%	64%	70%	71%	74%
Att 18	66%	61%	66%	60%	63%	76%	59%	66%	69%	67%
Att 19	70%	65%	73%	64%	65%	73%	62%	68%	67%	72%
Att 20	38%	39%	46%	39%	41%	61%	37%	42%	47%	55%

In addition, we have found that measures from both System 1 approaches tend to be moderately correlated with System 2 metrics such as brand appeal, likelihood to recommend, brand perceptions and other measures.

To date, however, we've never tested the IPT and EVT head-to-head to see if one approach performs better than the other. In addition, given the large number of screens that's required for the EVT (namely 42 screens when testing 28 emotions), we wanted to explore whether the IPT could be used to adapt the EVT. In this case, the EVT would become an adaptive, paired-comparison MaxDiff where we select 20 emotions from the initial IPT (we are essentially leveraging an express maxdiff with the initial IPT). The 20 emotions would be selected so that ~12 emotions are those that respondents agreed characterized the brand and where they responded quickly (i.e., they are emotions closely associated with the brand). The remaining ~8 emotions are those that are not associated with the brand (i.e., they disagree with) and that have longer response times. In this way, we are introducing greater differentiation into the adaptive EVT.

So to summarize, in this research we will compare:

- Implicit Priming Tests
- Emotional Valence Tests
- Adaptive Emotional Valence Test

METHODOLOGY

The study was fielded in the US through national panels. In order to get a good read on each of the System 1 approaches, respondents were randomly assigned one of the following survey paths:

1. IPT followed by Traditional EVT
2. Traditional EVT followed by IPT
3. IPT followed by Adaptive EVT

Groups 1 and 2 above were included to determine if there were any order effects (none were found during data analysis). In addition to a specific survey path, respondents were randomly assigned to evaluate one of four brands: Visa or Mastercard, both of which we hypothesized would evoke little or no emotional response (i.e., low emotional valence); or Fox News or MSNBC, which we consider to be high emotional valence brands given their polarizing nature. Panelists were also invited to complete the survey a second time about a week later (only a subset actually did so) and the data collected was used in the test-retest reliability assessment.

The following sample sizes were obtained for both the initial and post survey:

		Visa	Mastercard	Fox News	MSNBC News	Total
IPT then Traditional EVT	Initial Survey	50	51	57	49	207
	Post Survey	11	9	48	36	104
Traditional EVT then IPT	Initial Survey	50	53	49	50	202
	Post Survey	44	42	44	16	146
IPT -> Adaptive EVT	Initial Survey	99	104	98	99	400
	Post Survey	69	75	65	75	284
Total						809
						534

In total, we had a little over 800 completes from the initial survey with about 530 completes on the follow-up. So approximately 66% of respondents were a part of both phases of the study. For most of the analyses and reporting, we combined the samples of the low emotional valence and high emotional valence brand pairings (i.e., combined Visa and Mastercard, combined Fox News and MSNBC).

We manipulated the System 1 raw data to create the key measures used in the analysis. For IPT, we calculated the percentage of agreement for the positive and negative emotions across the low/high emotional valence brands. For the Traditional and Adaptive EVT, the average percentage of time that each emotion was selected as being associated with the brand tested was

calculated. Reaction time was also measured for the IPT exercise as there was no time limit in this exercise (for the EVT exercises they had to respond within two seconds).

For the outcome measures (i.e., System 2 evaluations), a number of different questions were included in the survey. One of those questions was a brand perception assessment based on nine measures (e.g., brand I most prefer, has an excellent reputation, best meets my needs, etc.). Through correlations and a factor analysis, we determined that for both the low and high emotional valence brand pairings, there was just one dimension; so, the nine measures were averaged. Other questions included likelihood to use/view in the next 12 months/7 days, likelihood to recommend, and proportion of spend on credit card/proportion of time viewing news network which were all used in our analysis.

In total, 28 emotions were tested in the study. These emotions come from a modified list of the Junto emotions.

Emotions Tested

Insecure	Content	Astonished
Anxious	Cheerful	
Frightened	Joyful	
Terrified	Euphoric	
Annoyed	Sentimental	
Frustrated	Affection	
Mad	Desire	
Enraged	Adoration	
Uncomfortable		
Appalled		
Repelled		
Revolted		
Unhappy		
Gloomy		
Depressed		
Despairing		
Confused		
Startled		
Stunned		

The emotions are displayed in 3 separate columns and are color-coded to indicate how they were categorized. The emotions displayed in red text are the negative emotions, those displayed in green text are the positive emotions, and the gray text references neither negative nor positive. These groupings were determined and confirmed through correlations and a factor analysis. The emotion “Astonished” loaded approximately equally on both the positive and negative factors, thus we separated it out rather than combining it with one of the factors. For some of our reporting, the average of the negative or positive emotions was used.

RESULTS

Correlation Analysis

The analysis begins with a review of the distribution of the means and standard deviations of each emotion with the brands across the three System 1 approaches. The results for IPT are displayed below:

IPT Distributions	Means				Standard Deviations			
	VISA	MASTERCARD	FOX NEWS	MSNBC NEWS	VISA	MASTERCARD	FOX NEWS	MSNBC NEWS
Insecure	14%	11%	37%	32%	0.34	0.31	0.48	0.47
Anxious	20%	13%	50%	37%	0.40	0.33	0.50	0.49
Frightened	12%	11%	40%	32%	0.32	0.31	0.49	0.47
Terrified	8%	9%	37%	30%	0.27	0.29	0.48	0.46
Annoyed	11%	11%	44%	39%	0.31	0.31	0.50	0.49
Frustrated	13%	10%	51%	44%	0.33	0.30	0.50	0.50
Mad	11%	9%	44%	37%	0.31	0.28	0.50	0.48
Enraged	14%	10%	41%	37%	0.35	0.30	0.49	0.49
Uncomfortable	11%	8%	46%	33%	0.31	0.27	0.50	0.47
Appalled	20%	15%	48%	47%	0.40	0.36	0.50	0.50
Repelled	11%	10%	40%	33%	0.31	0.30	0.49	0.47
Revolted	12%	11%	43%	34%	0.32	0.31	0.50	0.48
Unhappy	11%	11%	43%	36%	0.31	0.31	0.50	0.48
Gloomy	13%	10%	44%	40%	0.34	0.30	0.50	0.49
Depressed	9%	8%	37%	33%	0.28	0.27	0.48	0.47
Despairing	9%	11%	43%	41%	0.29	0.31	0.50	0.49
Confused	11%	7%	41%	32%	0.31	0.26	0.49	0.47
Startled	14%	14%	49%	38%	0.34	0.35	0.50	0.49
Stunned	15%	15%	53%	42%	0.35	0.36	0.50	0.50
Astonished	33%	25%	56%	51%	0.47	0.44	0.50	0.50
Content	86%	83%	59%	66%	0.35	0.38	0.49	0.48
Cheerful	80%	72%	47%	52%	0.40	0.45	0.50	0.50
Joyful	79%	70%	45%	49%	0.41	0.46	0.50	0.50
Euphoric	50%	33%	25%	30%	0.50	0.47	0.44	0.46
Sentimental	56%	47%	44%	41%	0.50	0.50	0.50	0.49
Affection	68%	55%	45%	48%	0.47	0.50	0.50	0.50
Desire	66%	67%	42%	48%	0.47	0.47	0.49	0.50
Adoration	56%	45%	39%	40%	0.50	0.50	0.49	0.49
Negative emotions	12%	11%	44%	37%	0.33	0.31	0.49	0.48
Positive emotions	68%	59%	43%	47%	0.45	0.47	0.49	0.49

The percentages on the Means table represent the proportions of respondents who agreed that said emotion is felt about said brand in the IPT exercise. For instance, 14% of respondents who answered for Visa agreed they feel insecure about the Visa brand. These values could range from 0% to 100%.

When we focus on the low emotional valence brands (Visa and MasterCard), we note a significant difference in the proportions of the negative emotions and the proportions of the positive emotions. In comparison, there is a pretty even distribution across the emotions on the high emotional valence brands (Fox News and MSNBC News). Similar patterns are noted on the standard deviations; lower standard deviations indicate a strong agreement on which emotions are associated with the brands. When we take a step back from the data and reflect, the results make sense. We do not expect consumers to have strong emotions associated with credit card brands, especially those that are negative, so there's less variability in the emotions associated with the brand. Generally, consumers have consistent positive/neutral emotions associated with credit card brands.

Similar to the results of the low emotional valence brands, the results of the high emotional valence brands are logical. The news networks, Fox News and MSNBC News, are very polarizing brands. Consequently, an even distribution of emotions occurs because some respondents love/like one brand, while disliking the other. Thus, as expected, the standard deviations are higher for the news network brands since consumers have very strong emotions associated with these brands and the feelings/attitudes towards these brands are not mutual.

tEVT Distributions	Means				Standard Deviations			
	VISA	MASTERCARD	FOX NEWS	MSNBC NEWS	VISA	MASTERCARD	FOX NEWS	MSNBC NEWS
Insecure	2%	3%	3%	4%	0.02	0.02	0.02	0.02
Anxious	4%	4%	5%	4%	0.05	0.02	0.10	0.02
Frightened	2%	2%	3%	3%	0.02	0.02	0.02	0.02
Terrified	2%	2%	3%	3%	0.02	0.02	0.02	0.02
Annoyed	2%	3%	4%	4%	0.02	0.02	0.03	0.02
Frustrated	3%	2%	4%	4%	0.02	0.02	0.02	0.02
Mad	2%	2%	3%	3%	0.02	0.02	0.02	0.03
Enraged	2%	2%	3%	3%	0.02	0.02	0.03	0.02
Uncomfortable	3%	3%	4%	4%	0.02	0.02	0.03	0.03
Appalled	3%	3%	4%	4%	0.02	0.02	0.02	0.02
Repelled	2%	3%	3%	3%	0.02	0.04	0.03	0.03
Revolted	2%	2%	3%	3%	0.02	0.02	0.03	0.03
Unhappy	2%	2%	3%	3%	0.02	0.02	0.03	0.03
Gloomy	3%	3%	3%	4%	0.02	0.02	0.02	0.02
Depressed	2%	2%	3%	4%	0.02	0.02	0.03	0.03
Despairing	3%	3%	3%	3%	0.02	0.02	0.02	0.02
Confused	3%	3%	3%	4%	0.02	0.02	0.02	0.02
Startled	4%	4%	4%	4%	0.03	0.03	0.02	0.02
Stunned	3%	3%	4%	4%	0.02	0.02	0.02	0.02
Astonished	4%	4%	4%	4%	0.03	0.02	0.03	0.02
Content	6%	6%	4%	4%	0.04	0.07	0.03	0.03
Cheerful	6%	7%	4%	3%	0.05	0.03	0.04	0.03
Joyful	7%	7%	4%	4%	0.06	0.04	0.04	0.03
Euphoric	5%	5%	3%	3%	0.05	0.03	0.04	0.03
Sentimental	5%	6%	4%	4%	0.04	0.04	0.04	0.03
Affection	6%	6%	4%	4%	0.04	0.03	0.04	0.03
Desire	5%	5%	4%	3%	0.03	0.03	0.03	0.03
Adoration	5%	5%	3%	3%	0.03	0.03	0.03	0.03
Negative emotions	3%	3%	3%	4%	0.02	0.02	0.03	0.02
Positive emotions	6%	6%	4%	4%	0.04	0.04	0.04	0.03

The results of the traditional EVT are evaluated next:

The percentages on the Means table represent the proportion of times an emotion was selected as being associated with the brand. For instance, across the 42 screens of the exercise, insecure was selected as an emotion associated with Visa 2% of the time at the aggregate level. These values could range from 0% to 7% as each item could appear up to 3 times across the 42 screens.

Less differentiation is seen across the emotions on the tEVT distribution compared to the IPT distribution, however, the patterns still hold. There's a flat distribution across emotions on Fox News and MSNBC News versus the difference in proportions between the positive and negative emotions on Visa and MasterCard.

As expected, the adaptive EVT distributions are similar. While we still note less differentiation compared to the IPT results, we notice a little more differentiation on the aEVT distribution compared to the tEVT:

aEVT Distributions	Means				Standard Deviations			
	VISA	MASTERCARD	FOX NEWS	MSNBC NEWS	VISA	MASTERCARD	FOX NEWS	MSNBC NEWS
Insecure	3%	3%	3%	3%	0.03	0.03	0.03	0.04
Anxious	3%	4%	3%	4%	0.04	0.04	0.04	0.04
Frightened	1%	2%	3%	3%	0.02	0.03	0.04	0.04
Terrified	2%	2%	3%	3%	0.03	0.03	0.03	0.03
Annoyed	2%	3%	4%	4%	0.03	0.03	0.04	0.03
Frustrated	2%	2%	4%	4%	0.03	0.03	0.04	0.04
Mad	1%	2%	4%	3%	0.02	0.03	0.04	0.03
Enraged	1%	2%	4%	2%	0.02	0.03	0.04	0.03
Uncomfortable	3%	3%	4%	3%	0.04	0.04	0.04	0.04
Appalled	3%	3%	4%	3%	0.03	0.03	0.04	0.04
Repelled	2%	2%	4%	3%	0.03	0.03	0.04	0.04
Revolted	2%	2%	4%	4%	0.03	0.03	0.04	0.04
Unhappy	2%	2%	4%	3%	0.03	0.03	0.04	0.03
Gloomy	2%	3%	3%	3%	0.03	0.04	0.04	0.03
Depressed	2%	2%	3%	4%	0.03	0.03	0.04	0.03
Despairing	3%	2%	3%	3%	0.04	0.03	0.04	0.03
Confused	2%	3%	3%	4%	0.04	0.03	0.04	0.03
Startled	3%	4%	3%	3%	0.03	0.04	0.04	0.03
Stunned	3%	3%	5%	3%	0.03	0.04	0.04	0.04
Astonished	4%	4%	4%	4%	0.05	0.04	0.04	0.04
Content	8%	9%	5%	6%	0.06	0.10	0.05	0.04
Cheerful	8%	8%	4%	5%	0.05	0.06	0.05	0.05
Joyful	9%	8%	4%	4%	0.06	0.06	0.05	0.04
Euphoric	5%	4%	2%	3%	0.05	0.04	0.03	0.04
Sentimental	6%	4%	4%	4%	0.05	0.04	0.04	0.04
Affection	7%	5%	4%	4%	0.05	0.05	0.04	0.04
Desire	6%	7%	3%	5%	0.05	0.06	0.04	0.04
Adoration	6%	4%	4%	4%	0.05	0.04	0.04	0.04
Negative emotions	2%	3%	3%	3%	0.03	0.03	0.04	0.04
Positive emotions	7%	6%	4%	4%	0.05	0.06	0.04	0.04

The percentages on the Means table represent the proportion of times an emotion was selected as being associated with the brand. For instance, across the 30 screens, insecure was chosen as an emotion associated with Vias 3% of the time at the aggregate level. These values could range from 0% to 10% as each item could appear up to 3 times across the 30 screens.

In summary, from the most to least differentiation across emotions, IPT comes out as our winner followed by aEVT and then tEVT.

Some may argue that it's unfair to compare the distributions of the three approaches since the mean proportions represent slightly different metrics. However, at the bottom of each distribution table, we've taken the average proportion for the negative and positive emotions. These metrics allow us to compare the results more confidently across the approaches. For IPT, the average proportion for positive emotions is about 5–6 times higher than the average proportion for negative emotions. In comparison, the average proportion for positive emotions is 2 times higher for tEVT and 3–4 times higher for aEVT. We also looked at the distributions re-proportioned on a 0 to 100 scale similar to the set-up of the IPT analysis. We noted consistently low proportions on the negative emotions for Visa/Mastercard on IPT while it was a bit more mixed (high and low proportions) across the negative emotions for tEVT and aEVT—again confirming our conclusions that the IPT provided more differentiation with its more extreme difference in results between the positive and negative emotion proportions.

In addition to evaluating the results separately by approach, we ran correlations to better understand the interrelationships between the three different System 1 approaches.

Correlations of Aggregate Results

Negative Emotions	VISA/MASTERCARD		FOX NEWS/MSNBC	
	IPT	Traditional EVT	IPT	Traditional EVT
	Traditional EVT	0.64	0.61	
Adaptive EVT	0.56	0.88	0.50	0.41

Positive Emotions	VISA/MASTERCARD		FOX NEWS/MSNBC	
	IPT	Traditional EVT	IPT	Traditional EVT
	Traditional EVT	0.83	0.87	
Adaptive EVT	0.97	0.90	0.97	0.78

When running correlations, we rolled up the emotions within the negative and positive dimensions to simplify the interpretation of the results. The low and high emotional valence brand pairings were also rolled up. The aggregated correlation analysis revealed quite strong relationships between the approaches, especially for the positive emotion associations. There are slightly less strong correlations between IPT and tEVT/aEVT for the negative emotions.

Correlations of Individual-Level Results

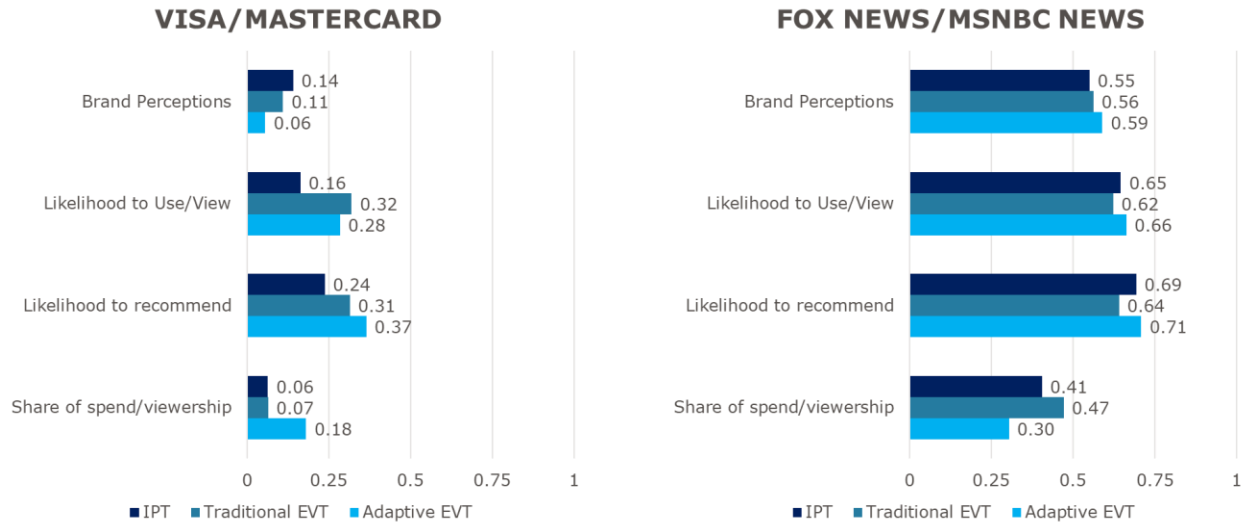
Negative Emotions	VISA/MASTERCARD		FOX NEWS/MSNBC	
	IPT	Traditional EVT	IPT	Traditional EVT
	Traditional EVT	0.43	0.69	
Adaptive EVT	0.36	-	0.71	-

Positive Emotions	VISA/MASTERCARD		FOX NEWS/MSNBC	
	IPT	Traditional EVT	IPT	Traditional EVT
	Traditional EVT	0.23	0.75	
Adaptive EVT	0.40	-	0.75	-

We also looked at correlations among the measures at the individual level. When reviewing these correlations, we still note relatively strong correlations between the approaches for the high emotional valence brands. However, with low emotional valence brands, the correlations reveal weak relationships between the System 1 approaches for Visa/MasterCard.

When considering future research with System 1 questions, it's important to consider that different System 1 measurement approaches will yield different results when looking at the results at the individual level when brands do not evoke strong emotional reactions. However, if doing research on low emotional valence brands and looking at the results in aggregate, the approach is less likely to have an impact.

After evaluating each of the System 1 approaches separately then together, we looked to understand the relationship between these approaches and the System 2 metrics tested. The System 2 metrics tested included brand perceptions, likelihood to use/view, likelihood to recommend, and share of spend/viewership.



These correlations are calculated by averaging the absolute values of the positive and negative emotion variable correlations with the attitudinal/behavioral metrics. The negative and positive variables were created by rolling up the proportions of emotions associated with brand.

Based on these correlations, we concluded there is no single approach that stands out or is consistently high across the System 2 metrics. In terms of the relationship between the System 1 and 2 metrics, we found these to be relatively weak for the low emotional valence brands, but relatively strong relationships for the high emotional valence brands. This suggests, particularly with brands that are less likely to evoke strong emotion, that System 1 and System 2 metrics provide different insights and both contribute to understanding perceptions of brands.

Regression Modeling

We also ran regression models to understand the impact the various System 1 approaches have on the System 2 likelihood to use/likelihood to view metric. We analyzed models which replicated the survey paths (i.e., some respondents went through the IPT and traditional EVT exercise while others went through the IPT and adaptive EVT). First, we look at the credit card models.

The tables below show the models for respondents who went through the IPT and Traditional EVT and then for those who completed the IPT and Adaptive EVT. There are a few things to note:

1. The R-square values for both analyses are very weak ($\sim .10$). This suggests these variables do little to explain likelihood to use.
2. In both models, the reaction time variable (IPT Speed Index) has virtually no impact as evidenced by the very small coefficients.
3. Of the remaining variables, few have significant coefficients, and in the case of the IPT/aEVT model, we see a reversal with respect to the aEVT positive emotions (it's negative, but we would expect it to be positive). The reason for this reversal is that multicollinearity played a role among the variables.

Credit Card Regression Models

Adj R2:	IPT + Traditional EVT		Beta Coefficient	Significance	0.113
	IPT Negative Emotions		-0.021	0.148	
	IPT Positive Emotions		0.106	<.001	
	Traditional EVT Negative Emotions		-0.089	0.352	
	Traditional EVT Positive Emotions		0.199	0.036	
	IPT Speed Index		0.001	0.914	

Adj R2:	IPT + Adaptive EVT		Beta Coefficient	Significance	0.100
	IPT Negative Emotions		-0.122	<.001	
	IPT Positive Emotions		0.099	<.001	
	Adaptive EVT Negative Emotions		-0.551	<.001	
	Adaptive EVT Positive Emotions*		-0.353	<.001	
	IPT Speed Index		-0.022	0.095	

In contrast, the tables below show the regression results for the high emotional valence TV network brands. In this case, we find that the models have much stronger R-square values meaning that the System 1 variables do a better job of explaining the variability in likelihood to watch. This aligns with the results of the correlations shown earlier that demonstrate the brands that evoke stronger emotions are more closely related to System 1 emotional assessments than those for which respondents feel a weaker emotional response. However, like the credit card regressions, we note that reaction time (IPT Speed Index) has no impact and we again see the counterintuitive negative coefficients for positive emotions, which can be attributed to multicollinearity. The IPT measures, on the other hand, seem to perform very well in both of the regressions below.

News Network Regression Models

Adj R2:	IPT + Traditional EVT		Beta Coefficient	Significance	0.532
	IPT Negative Emotions		-0.213	<.001	
	IPT Positive Emotions		0.458	<.001	
	Traditional EVT Negative Emotions		-0.57	<.001	
	Traditional EVT Positive Emotions		-0.435	<.001	
	IPT Speed Index		0.007	0.454	

Adj R2:	IPT + Adaptive EVT		Beta Coefficient	Significance	0.600
	IPT Negative Emotions		-0.384	<.001	
	IPT Positive Emotions		0.383	<.001	
	Adaptive EVT Negative Emotions		-0.685	<.001	
	Adaptive EVT Positive Emotions*		-0.575	<.001	
	IPT Speed Index		0	0.963	

Reliability Test

Finally, we wanted to look at the results over time to assess the reliability of the System 1 measures. As described previously, a subset of respondents completed the survey two different times, about a week apart. With this data, we were able to look at the correlations between their System 1 and System 2 responses.

The table below shows there are strong correlations for both the System 2 (i.e., likelihood to use/view, likelihood to recommend and share of spend/viewership) and System 1 (i.e., IPT, tEVT, and aEVT) measures for the credit cards brands along with the TV networks. We do see that the lower emotional valence credit card brands have weaker relationships/correlations compared to the TV networks, but they are still acceptable. The one significant difference is among the TV networks, the correlation for tEVT is significantly higher than that of IPT and aEVT (but both of the latter measures still perform very well). Particularly encouraging is that the System 1 measures have approximately the same reliability as traditional System 2 measures. This suggests that researchers can have confidence that they will receive stable results over time when using any of these System 1 measures.

**Correlations
(Initial Survey with Follow-Up)**

	VISA/ MASTERCARD	FOX NEWS/ MSNBC NEWS
Likelihood to Use/View	0.60	0.89
Likelihood to recommend	0.61	0.89
Share of spend/viewership	0.89	0.78
IPT	0.61	0.81
Traditional EVT	0.69	0.89
Adaptative EVT	0.62	0.79

DISCUSSION

The table below summarizes the results at a high level.

Analyses	Visa/Mastercard			Fox News/MSNBC		
	IPT	Trad. EVT	Adapt. EVT	IPT	Trad. EVT	Adapt. EVT
Distributions	↑	-	↑	↑	-	↑
Perceptions/ Intentions	↓	↓	↓	↑	↑	↑
Regressions	↓	↓	↓	↑	↓	↓
Test/retest	↑ -	↑	↑	↑	↑ +	↑

Across the low and high emotional valence brands, the IPT approach appears to have provided a little more differentiation across the brands, followed by adaptive EVT. But at the end of the day, they all demonstrated good differentiation.

When we looked at the relationships between the System 1 approaches and respondents' perceptions and intentions, we found much higher correlations across the high emotional valence brands compared to our low emotional valence brands. Within low and high emotional valence brands there were no consistent differences between the three System 1 approaches, all three performed in a roughly similar fashion.

When looking at the regression models, the IPT in particular had the best relationship between System 1 and System 2 metrics for high emotional valence brands. In addition, we found that the reaction time within the IPT approach does not appear to be an important measure. Moving forward, we do not believe it is important to include reaction time when measuring System 1 via IPT.

And finally, when examining the test-retest results, we found that all brands and approaches exhibited very high correlations across the board, suggesting that the three System 1 approaches are good, reliable measures.

So, considering all of our findings, we feel that all three approaches work very well and do a good job of assessing System 1 processing. However, given that the IPT approach is very easy to design, easy for respondents to answer and doesn't involve creating an experimental design, we feel that IPT really warrants consideration in all aspects compared to traditional EVT and adaptive EVT. The latter two approaches work well, but do involve more effort in design and analysis, and take respondents longer to complete.



Michael Patterson



Sonia Hundal

REFERENCES

- Kahneman, D. (2011) Thinking, Fast and Slow. London: Penguin.
- Lipovetsky, S. (2020) Express analysis for prioritization: Best-Worst Scaling alteration to System 1, *Journal of Management Analytics*, 7:1, 12–27.
- Patterson, M., Frazier, C (2017) Choice Models vs. Implicit Association Tests: Which Assess Brand Preference More Accurately? Presentation at Sawtooth Software/SKIM conference.

ADAPTIVE CONJOINT: TESTING BEST PRACTICES AND METHODS

ZACHARY LEVINE
KEES VAN DER WAGT
SKIM

BACKGROUND

Choice-based conjoint, often abbreviated as CBC, has been a tried-and-true method of conducting market research within multi-attribute categories for decades. Designs for CBC surveys are usually carried out with an emphasis on balance; to wit, each level within each attribute should appear with roughly equal frequency (one-way balance), and each combination of levels within each possible pair of attributes should also appear equally often (two-way balance). However, we can challenge the assumption that optimizing for design is best practice. We observe that consumers typically make decisions based on one or two key attributes rather than a holistic accounting for every possible attribute; for instance, some consumers are very brand loyal, while others may pick whichever product has the lowest price. Therefore, rather than always aiming for balance, there is potential to gain greater insight and granularity for all attributes for a given consumer by aiming to emphasize each respondent's most preferred levels of each attribute.

Of course, this has been tried before, not only in previous research but also in Sawtooth Software's Adaptive Choice-Based Conjoint module, often abbreviated as ACBC. Within ACBC, respondents do see differing level frequencies, depending on their responses to a prior Build-Your-Own (BYO) task. SKIM has presented research on ACBC several times, such as in the paper from Hoogerbrugge, Hardon, and Fotenos at the 2013 Sawtooth Software Conference, which (among other findings) found that ACBC significantly outperformed CBC in all variations of ACBC tested. However, ACBC also has its imperfections, as outlined in the paper from Hoogerbrugge and Hardon at the 2018 Sawtooth Software Conference:

- ACBC can be too extreme; the BYO task for each respondent is based on their ideal levels within each attribute, but respondents often have 2–3 preferred levels for a given attribute instead. Because of this, trade-offs between multiple preferred levels may not get captured as easily.
- ACBC assumes for *all* attributes that there is a level that is dominant over other levels, again by nature of the setup of the BYO exercise. However, this assumption may not be true. As mentioned above, there are very often one or two attributes which especially drive respondent choice, such as brand or price. Instead of emphasizing preferred levels for all attributes, it may make more sense to emphasize preferred levels just for those most important attributes, and show levels with roughly even frequency for less important attributes.

Due to these factors, researchers at SKIM have developed and refined a method of adaptive conjoint called *preference-based conjoint*, which we will abbreviate as *PBC*. PBC does not make use of a BYO task. Instead, it shows each level of each attribute to a respondent in proportion to frequency of choice in that respondent's prior choice tasks. In constructing our method this way,

we are able to account for the possibilities of a respondent having 2-3 preferred levels within a given attribute rather than just one, and also allow for a more even preference structure for some attributes than for others. In short, this method has maximum flexibility.

In addition to PBC, SKIM's Kees van der Wagt has in the past year created an entirely new method of adaptive conjoint, referred to in this paper as *on-the-fly latent class*, which we can abbreviate as OLC. We will expand on this method in detail, but the spirit of this method is similar to PBC (and ACBC): it aims to emphasize the most relevant concepts for each respondent. However, OLC differs from PBC in that it evaluates relevance and selects *entire concepts* rather than selecting preferred levels *within attributes*, and it uses previous respondents' data to properly determine latent class segments.

PREVIOUS RESEARCH ON PBC

This paper takes inspiration from and is meant to elaborate on prior research from the aforementioned 2018 paper by SKIM's Jeroen Hardon and Marco Hoogerbrugge. In that paper, the authors laid out two possible approaches to preference-based conjoint:

1. An approach based on ACBC, where the most preferred level itself is given somewhat less frequency than in standard ACBC (but still highest), levels adjacent to the most preferred level are given considerably increased frequencies, and the other levels far away from the preferred level remain at low frequencies.
2. A new approach that is not based on ACBC, but directly uses the levels chosen in previous tasks as possible levels in subsequent tasks. So, if there are six levels of a given attribute, in task 4 (for instance) there would in effect be *nine* levels to choose from: the six original levels, plus three flexible levels that would take the form of the levels of the chosen concepts in tasks 1-3. So, these chosen levels would be more likely to show in task 4 than other levels. This would also mean that in later tasks, the on-the-fly nature of the survey would increase.

This paper used two test studies in the mobile telephone space, and used a very simplified holdout task with dozens of possible responses, as one aim of the study was to test the effectiveness of different approaches to PBC on a potential simulator with many products included.

Approach #2 outperformed approach #1 in both test studies, as well as standard implementations of ACBC and CBC. In addition, approach #2 was found to be more flexible and better individualized than approach #1, as one disadvantage of ACBC remains in approach #1: all attributes are forced to adhere to a certain probability distribution, regardless of their importance to any given respondent, whereas in approach #2 the probability distribution for a given attribute is tangibly changed for each respondent only if it becomes clear that it is a relevant choice criterion for that respondent.

Because of the above factors, we went with an approach based on approach #2 in our study for this conference.

PBC SETUP

We wish to keep the flexibility of approach #2 in our PBC approach for this study, and we take a similar approach to decide which level within any given attribute is shown in each task. This is the very general idea:

1. At the start of the survey, each level is equally likely to be shown.
2. For each task, we add an additional flexible level corresponding to the level of the chosen concept in each previous task.

This directly mirrors approach #2 taken in the Hardon/Hoogerbrugge study from 2018. However, there are still assumptions inherent in this approach. For one, is it necessary to strictly add *one* flexible level corresponding to chosen concepts in prior tasks? Adding more than one would allow us to converge on preferred levels more quickly, which may be more desirable; it could be that some attributes, like brand, exhibit very clear preference and would be well suited to allow for quicker convergence, so that we may for instance show a respondent's top brand levels more often in later tasks once we already have an idea of their true preference structure within brand. Also, we may wish to add more flexible levels within attributes like brand or price where we assume that preferences are clearer and more likely to drive respondent choice than in attributes that may be weaker drivers of choice. To this end, we develop a system of developing a probability structure within each attribute in any given task, as well as a new term that we have internally referred to as *boost weight*.

1. Before task 1, each level within any given attribute has a score of 1.
2. We define before the survey a *boost weight* for each attribute: this essentially corresponds to the number of flexible levels that we assign to the levels within this attribute that appear in each chosen concept. A higher boost weight for an attribute leads to quicker convergence within that attribute.
3. Assuming we are showing task N, we assign each level a score of $1 + (\text{boost weight}) \times (\text{number of concepts with this level chosen})$.
4. The probability of any given level showing in a concept is proportional to its score. For each concept, we generate a random number between 0 and 1 and determine which level is shown in this attribute from this random number.

We can more clearly show this through an example, where we conceive an attribute, *color*, with four levels: *red*, *blue*, *green*, and *yellow*. Let us also assume in this example that the first four tasks of this survey are standard CBC tasks, and that the tasks that are dynamically generated through PBC begin in task 5. Finally, we assign the color attribute a boost weight of 2.

- The levels chosen within our color attribute in the first four tasks are as follows: blue, blue, green, yellow. Although we do not dynamically generate tasks 1-4, we still keep track of what is chosen in these tasks.
- This means that after 4 tasks, blue has been chosen twice, green and yellow once each, and red not at all. So after 4 tasks, we assign blue a score of $1 + (2 \times 2) = 5$, green and yellow a score of $1 + 2 = 3$, and red a score of 1.
- Accordingly, the probability for each concept in task 5 that each level will be drawn in our color attribute is as follows: red has a probability of $1/12$, blue has $5/12$, and green and yellow each have $1/4$. For each concept, if our random number draw is between 0 and

~.08333, we assign that concept red; if between ~.08333 and .5, we assign blue; if between .5 and .75, green; if greater than .75, yellow.

- We draw .84, .46, .27, .65, so our concepts in task 5 get levels [yellow, blue, blue, green] within the color attribute.

This general method is the basis behind all tasks that are dynamically generated through PBC. In practice, like in the 2018 study, we tend to move away from an even probability structure and closer to a distribution with emphasis on preferred levels as we get later on in the survey.

PBC TEST LEGS

In addition to further verifying the effectiveness of our PBC approach in comparison to CBC, we wished to test a few variations within PBC:

- We tested four different possibilities for boost weight: 1 for all attributes, 2 for all attributes, 5 for all attributes, and a more flexible approach where brand/price get 4 while others get 1 or 2. This spans a range from slower to quicker convergence, and covers the plausible scenario that brand and price are key decision drivers, in which case it would make sense to allow these to converge more quickly to allow for some more granularity in other attributes.
- We tested both with and without a partial-profile “warm-up” with three tasks, which did not factor into any PBC calculations but allowed for the respondent to settle into favorite brands, tier, or practice trade-offs.
- We also tested whether it is better to start dynamically generating tasks at task 2 (earlier) or task 5 (later), with any tasks beforehand as standard CBC. A later start would cause us to wait until task 5 to converge around preferred levels, allowing us to begin moving toward preferred levels only when we have a bit more information.

VAN DER WAGT’S ON-THE-FLY LATENT CLASS: BACKGROUND AND SETUP

In 2022, Kees van der Wagt developed an alternative approach to adaptive conjoint that takes into account preferences from other respondents as well as previous choices within the survey. Unlike PBC, it chooses what to show on each screen on the entire concept level rather than determining within each attribute one at a time. The general idea is this:

- The overall goal is to show the most relevant concepts to any given respondent during the survey.
- We can first run latent class on previous data and use the first few tasks to determine a respondent’s LC segment. We map all possible concepts in our design space on these LC segments as a first read on what would be most relevant to our respondent.
- As we get more data from the current respondent, we can use prior choices in addition to previous respondents’ LC classes to help us decide what will be most relevant for the respondent to see. The further along in the survey, the more the determination of what to show is based on respondent data rather than other respondents’ LC segments.
- The way we determine what is most relevant for the respondent: we calculate one- and two-way frequencies from the top concepts, where “top” is determined by on-the-fly

utility calculation. We then pick concepts that get us the closest to those desired frequencies.

A potential benefit of this approach is that we can make use of data from previous respondents; we can make a baseline assumption based on actual prior data rather than the baseline assumption in PBC that all preferences within each attribute are equal. The specific approach to on-the-fly latent class (OLC) we took in this study would read as follows:

1. Use 1/3rd of all tasks to identify the respondent's LC segment. After each task, the best next task is calculated. (Best as in "best capable of predicting the LC segment.")
2. Given the respondent's answers and the LC segments (based on those answers), calculate "respondent utilities," using a weighted average for respondent's answers and the LC segments.
3. Calculate the utility for all concepts in the respondent's design version AND the next 2-4 design versions. (The respondent's design version is determined within the survey.)
4. Calculate the one- and two-way frequencies from the top N% of concepts. N could theoretically be higher, lower, or adaptive; for simplicity's sake, we tested top 50% and top 20% in this study.
5. Check across all concepts from step 3 which concept to add to get closest to the desired one- and two-way frequencies from step 4. All two-ways together have the same weight as all one-ways together.

TEST STUDY

We conducted a test study of mobile telephone subscriptions in the US market in early 2023. We had 10 tasks, with 4 concepts per task. We used 13 total attributes with the same levels across all legs of this study:

1. Brand (12 levels)
2. High speed data (9 levels)
3. Data carry-over (5 levels)
4. Mobile hot spot (8 levels)
5. Included video service (10 levels)
6. Included TV service (8 levels)
7. Video streaming resolution (3 levels)
8. Phone theft + loss protection (2 levels, yes/no)
9. Accidental phone damage protection (2 levels, yes/no)
10. Cloud storage (6 levels)
11. International calling (7 levels)
12. Security and monitoring (7 levels)
13. Price (8 levels, \$10–\$100)

Our partial profile warm-up contained brand, high speed data, data carry-over, mobile hot spot, video service, TV service, video streaming resolution, and price, with the same levels as in the main exercise. 3 tasks, 4 concepts per task.

We also included a standard CBC leg, with design generated through balanced overlap, to have a basis of comparison, with design specifications the same as for PBC and OLC.

Sample sizes: N = 993 for PBC, N = 536 for standard CBC, N = 337 for OLC. Sample distributed evenly across legs within PBC and OLC.

HOLDOUT TASKS

We used three holdout tasks in this study, all with a range of prices included. The specifications of each holdout task were as follows:

1. An unbranded holdout task with only three concepts and all other attributes included. 3 total concepts.
2. An unbranded partial-profile holdout task with the following attributes: high-speed data, video service, TV service, video streaming resolution, phone theft + loss protection, and price. 9 total concepts.
3. A branded partial profile task with all of the attributes in task #2 plus brand. 9 total concepts.

RESULTS

We evaluate this study based on hit rate and mean absolute error (MAE) in our three holdout tasks, each of which we evaluate separately. We constrained our price attribute to have decreasing utility as price increased, but applied no other constraints in estimation. We used Sawtooth Software's CBC/HB tool with 50,000 total iterations (30k burn-in + 20k used).

Overall, this study corroborates the finding from the 2018 Hardon/Hoogerbrugge study that PBC performs as well as or better than standard CBC. The 2018 study was more positive on PBC's overall performance vs. CBC than this one, but PBC still has even performance vs. CBC with a significantly better result in hit rate for the partial profile unbranded task.

Van der Wagt's on-the-fly latent class (OLC) method actually shows significantly lower MAE in the 3-concept unbranded full-profile holdout task than either PBC or standard CBC, suggesting potential in smaller simulation spaces. However, it performs worse than PBC and CBC in both hit rate and MAE in the 9-concept full-profile holdouts.

Hit Rates: Overall by Methodology

	Holdout 1 (full profile unbranded, 3 concepts)	Holdout 2 (partial profile unbranded, 9 concepts)	Holdout 3 (partial profile branded, 9 concepts)
PBC	49.9%	45.8%	50.7%
CBC	50.0%	37.9%	50.9%
Van der Wagt's OLC	47.8%	33.8%	27.6%

MAE: Overall by Methodology

	Holdout 1 (full profile unbranded, 3 concepts)	Holdout 2 (partial profile unbranded, 9 concepts)	Holdout 3 (partial profile branded, 9 concepts)
PBC	8.2%	2.5%	2.8%
CBC	8.0%	2.6%	2.8%
Van der Wagt's OLC	1.5%	7.7%	8.6%

Within Van der Wagt's OLC, we do see some improvement in partial-profile holdouts when we use the top 20% of concepts as “most relevant,” rather than using the top 50%. However, this is not enough to make up the gap with PBC and CBC.

Hit Rates: Using Top 20% of Concepts in OLC vs. Top 50%

	Holdout 1 (full profile unbranded, 3 concepts)	Holdout 2 (partial profile unbranded, 9 concepts)	Holdout 3 (partial profile branded, 9 concepts)
20 percent	46.8%	35.3%	30.1%
50 percent	46.8%	32.3%	46.8%

MAE: Using Top 20% of Concepts in OLC vs. Top 50%

	Holdout 1 (full profile unbranded, 3 concepts)	Holdout 2 (partial profile unbranded, 9 concepts)	Holdout 3 (partial profile branded, 9 concepts)
20 percent	1.9%	6.9%	8.3%
50 percent	1.3%	8.9%	9.5%

Our most clear finding within PBC is that adding our partial-profile warm-up helps modestly with both hit rate and MAE across all holdout tasks. This is not terribly surprising, as intuitively allowing respondents to take a few questions to align on their own preferences within the survey can only help.

Hit Rates: Partial Profile On/Off Within PBC

	Holdout 1 (full profile unbranded, 3 concepts)	Holdout 2 (partial profile unbranded, 9 concepts)	Holdout 3 (partial profile branded, 9 concepts)
With Partial Profile	50.3%	46.2%	51.0%
Without Partial Profile	49.6%	45.5%	50.4%

MAE: Partial Profile On/Off Within PBC

	Holdout 1 (full profile unbranded, 3 concepts)	Holdout 2 (partial profile unbranded, 9 concepts)	Holdout 3 (partial profile branded, 9 concepts)
With Partial Profile	6.5%	2.5%	2.6%
Without Partial Profile	9.7%	2.7%	3.0%

In determining whether it is better to begin dynamic generation in task 2 or task 5, the results are more ambiguous and differ across holdout tasks. Still, in two out of the three holdout tasks each for both hit rate and MAE, starting dynamic task generation in task 2 performs best.

Hit Rates: Beginning Dynamic Generation in Task 2 vs. Task 5

	Holdout 1 (full profile unbranded, 3 concepts)	Holdout 2 (partial profile unbranded, 9 concepts)	Holdout 3 (partial profile branded, 9 concepts)
Task 2	49.1%	46.6%	52.0%
Task 5	50.8%	45.0%	49.2%

MAE: Beginning Dynamic Generation in Task 2 vs. Task 5

	Holdout 1 (full profile unbranded, 3 concepts)	Holdout 2 (partial profile unbranded, 9 concepts)	Holdout 3 (partial profile branded, 9 concepts)
Task 2	7.5%	2.5%	3.0%
Task 5	9.0%	2.8%	2.6%

In our unbranded holdout tasks, we see uniform boosting generally outperform variable boosting by attribute in both hit rate and MAE. We see a slight trend toward quicker boosting, as a uniform boost weight of 5 has lowest MAE in the unbranded full-profile holdout (and is about even with weight 1 in the unbranded partial-profile). Meanwhile, a uniform weight of 2 performs best in hit rate in both unbranded tasks. However, in the branded holdout task, this changes; varying boost weight by attribute (and weighting brand/price more), which performs worst in the unbranded holdouts, has lowest MAE and is just behind uniform weight 1 for highest hit rate.

Hit Rates: Boost Weights of 1, 2, 5, and Variable by Attribute

	Holdout 1 (full profile unbranded, 3 concepts)	Holdout 2 (partial profile unbranded, 9 concepts)	Holdout 3 (partial profile branded, 9 concepts)
Uniform Weight 1	47.9%	44.6%	54.9%
Uniform Weight 2	53.9%	48.2%	49.6%
Uniform Weight 5	51.9%	45.4%	43.2%
Variable Wt By Attr.	46.7%	44.8%	53.0%

MAE: Boost Weights of 1, 2, 5, and Variable by Attribute

	Holdout 1 (full profile unbranded, 3 concepts)	Holdout 2 (partial profile unbranded, 9 concepts)	Holdout 3 (partial profile branded, 9 concepts)
Uniform Weight 1	6.9%	2.2%	2.7%
Uniform Weight 2	7.6%	2.8%	3.2%
Uniform Weight 5	6.1%	2.3%	4.1%
Variable Wt By Attr.	10.9%	3.3%	2.3%

CONCLUSIONS

As in the 2018 study, we observe that PBC is a worthwhile alternative to CBC for our simulators. In both studies, we are operating in a large design space in which respondents will generally make their decisions based on one or two attributes rather than evaluating them all equally. We hypothesized that a solution that allows for imbalanced level frequencies might allow us to get a good read on a respondent's "lesser" attributes, and this approach appears to work.

The least ambiguous finding in this study was the positive impact of a partial-profile warm-up; while the numbers themselves are not dramatic, the direction was positive within both hit rate and MAE, across all three holdout tasks. As the initial PBC tasks are very important in determining a respondent's "top few" preferences within each attribute, allowing for a few tasks not used in analysis or in the main PBC module that can help respondents calibrate preferences is a worthwhile exercise.

The soft trend toward starting dynamic generation earlier rather than later is slightly surprising; intuitively, it would make sense that a few more tasks to allow the respondent to calibrate their preferences would allow for more clarity in the dynamic PBC tasks later on. However, we observe that the beginning of a PBC module somewhat resembles a random CBC design anyway, with greater imbalance later in the study, so having more standard CBC tasks may indeed be unnecessary.

It makes sense that the branded holdout task would be the one where specifically emphasizing brand (and price) in our boost weights bears the most fruit. However, in actual simulators, brand is considered more often than not, which may help inform best practice in such cases.

We cannot yet conclude that OLC is clearly a worthy alternative to CBC (or PBC in the adaptive realm), but it is worth exploring more; this study was the first time it was tested in a research forum such as this one. A study dedicated mainly to it, with a larger sample size, more holdout tasks, and more variation within, could provide interesting results not seen here. One note: OLC chooses entire concepts to show next instead of assigning levels one attribute at a time. This may explain the disparity in OLC's performance between the full-profile and partial-profile holdouts; the partial-profile holdouts both have concepts that do not resemble what was tested in the main module and have more concepts (9) than were tested on screen (4), and it may be that given the nature of OLC that it is more sensitive to such a disparity than PBC.

FURTHER CONSIDERATIONS

At the A&I Summit in Barcelona, the point was made that even more than three holdout tasks would be best practice; the specific number given was *five*, with realistic level overlap. In particular, the Hardon/Hoogerbrugge study made use of holdout tasks that had far more than even nine concepts, but had a few dozen, very simplified, concepts. As conjoint simulators often occupy spaces that make room for several concepts, including a holdout task or two that would better reflect this reality would be a good idea in future research. In addition, we only included one branded holdout task; given the differences we see in the branded holdout task, there would perhaps be interesting consequences from the inclusion of more. The aforementioned large holdouts from the 2018 study were branded.

It would probably be worth rethinking the size of our design space. With 13 attributes and more than 4 levels in many of our attributes, this design space was simply very big, and research on a smaller design space might yield more precise results. In future research of this kind, we would consider down-scaling the main exercise as well as any partial-profile warm-ups or holdouts. Incidentally, this is another case for having larger holdout tasks—even in a smaller design space than this one, there are many, many possible products that may be included. Also, the effect of a summed-price attribute versus a conventional price attribute is worth exploring.

In testing a partial-profile warmup, it would be good in the future to vary the attributes shown. 10 tasks may be too few for this research; we used the first 4 tasks to determine LC segment in our on-the-fly latent class method, and we designated the first 4 tasks as standard CBC for half of respondents in the PBC leg. 12 tasks would have been preferable.

In future studies of this nature, we will consider out-of-sample holdout validation, as our MAE and hit rate calculations were based on in-sample holdout, which can reward overfitting to the idiosyncratic aspects of the sample used for both utility estimation and holdout validation—a point made by Bryan Orme during his discussant comments.



Zachary Levine



Kees Van der Wagt

BIBLIOGRAPHY

- Hardon, Jeroen and Hoogerbrugge, Marco, SKIM Group (2018). Preference-Based Conjoint: Can It Be Used to Model Markets with Many Dozens of Products? *Proceedings of the 2018 Sawtooth Software Conference*.
- Hardon, Jeroen, Hoogerbrugge, Marco, and Fotenos, Christopher, SKIM Group (2013). ACBC Revisited. *Proceedings of the 2013 Sawtooth Software Conference*.

ALTERNATIVE-SPECIFIC CONJOINT FOR PRODUCT DEVELOPMENT AND PRICING IN TECH AND DURABLES CATEGORIES

FAINA SHMULYIAN
TYLER DUGAN
BIG VILLAGE

ABSTRACT

Choice-Based Conjoint (CBC) analysis is widely used for multiple applications including feature, price, and assortment optimization. Its properties make it attractive in complex categories, such as tech and durable goods. But sometimes, especially in these categories, selecting the best design for a choice exercise might present a challenge for a researcher. This paper focuses on the alternative-specific design built from structural relationships between complex attributes. The paper compares an alternative-specific design with other common CBC designs often used in complex categories, including a traditional CBC, a shelf test, and an adaptive CBC. It considers the impact of a design choice on models, simulations, conclusions, and recommendations. The case study presented in the paper shows similarities and differences between CBC approaches providing practical hints for researchers using the methodology.

All considered models demonstrate sufficient accuracy of estimation and stability in presence of noise. The case study underlines differences in respondents' reactions to choice exercises based on various CBC designs. The alternative-specific CBC and the traditional CBC draw attention to particular product features and characteristics. The shelf test accurately evaluates a limited number of fixed configurations and configuration/price trade-offs. And the adaptive CBC studies configurations and prices that are the most relevant to each respondent but has limited ability to capture switching behavior.

INTRODUCTION

Choice-Based Conjoint (CBC) analysis or Discrete Choice Modeling (DCM) is one of the most popular and trusted tools for product development and price optimization in modern survey-based research. In particular, CBC is applied in complex categories like tech and durable goods. Often, it is used to test innovations in these categories. What makes conjoint so attractive for these kind of studies?

First, CBC mimics real-world purchasing scenarios and behaviors. A respondent considers products or offerings in a competitive context, evaluating different features and prices and making reasonable trade-offs. The data collected in a discrete choice exercise can be utilized in a Hierarchical Bayesian estimation to model preferences individually for each respondent and accurately describe market heterogeneity. Using CBC, researchers can simulate consumer behavior in hypothetical scenarios, testing products, offerings, and price points that don't exist on the market yet. CBC analysis can account for interactions between different product attributes, simulate various scenarios on the market, optimize product features in these scenarios, and estimate price sensitivity. The analysis can be used to identify thresholds and optimal prices. Relative importance of attributes can also be estimated in a CBC.

Another advantage of a CBC is its flexibility. Modern conjoint offers multiple design options to accommodate various business objectives related to different products and categories. It

allows describing not just a large number of attributes and levels, but also a complex layout with a hierarchy of attributes, when a level of a primary attribute defines other conditional attributes.

Sometimes, especially in complex tech or durables categories, products may have a unique set of attributes. As an example, let us consider headphones or earbuds. These products have common attributes—brand, shape, color, price, etc. Levels of these attributes would be presented in every alternative in a CBC. For example, earbuds can be wired or wireless. Also, there will be attributes/features that are presented only for wireless or wired earbuds. For wireless earbuds, these attributes could be charging time and battery life, and for wired earbuds it could be in-line volume control, plug type, etc. To model a category like headphones and earbuds using a CBC, many researchers are utilizing alternative-specific designs. With this kind of design, every primary attribute (such as wireless or wired earbuds in the example above) will only be paired with a relevant subset of conditional attributes (such as charging time and battery life for wireless earbuds). An alternative-specific design only considers product alternatives containing features that make sense for respondents in this category. Therefore, in an alternative-specific CBC, respondents are able to make meaningful tradeoffs and choose between products with feasible combinations of attributes. Other examples of appropriate categories to apply an alternative-specific CBC could be computers (desktops, laptops, tablets), electric floor-cleaning devices (vacuum cleaners, wet cleaners, wet-dry combos). Based on data collected in a choice experiment with an alternative-specific design, we will show that an accurate model can be built and used for feasible product development.

Overall, the alternative-specific approach makes CBC designs more flexible and “compact” in studies involving complex products or offerings. It provides a more realistic description of alternatives, allowing respondents to choose from the most relevant options, which improves data quality. Classifying and presenting attributes as common, primary, and conditional better informs the estimation and ensures higher accuracy of modeling in CBC studies.

As any other type of advanced CBC design, alternative-specific conjoint has its limitations. In general, an alternative-specific design can be more detailed than a standard CBC, but it still assumes a certain level of generalization in product descriptions. Introducing additional prohibitions or conditions is not recommended in an alternative-specific conjoint. Alternative-specific CBC is not suitable for estimating interactions since conditional attributes are directly associated with primary attributes in these type of studies. If almost every product in a category has its own set of attributes, a shelf test might be more appropriate for the study than an alternative-specific design. If products have a very large number of attributes, a partial profile approach or an adaptive CBC could be a better fit. Generally, we will characterize context that leads to appropriate designs for CBC.

PROBLEM STATEMENT

After learning all the details of a business problem from a client and deciding that a CBC would be a good fit for the study, a researcher has to choose the best design that meets the business objectives and accurately describes the situation of the market. Sometimes, the choice of a design is obvious, but if we are dealing with a complex category and product, it might not be so easy to select the best CBC approach. It could be an alternative-specific design which would fully recreate a layout with primary and conditional attributes. Or it could be a traditional design which would simplify the structure and help to implement complex relationships between the attributes with prohibitions and conditions; a traditional CBC would encourage a greater focus on a small number of attributes. A shelf test would be a trivial structure with only two attributes:

product and price. It ensures in-depth testing of these fixed products, assortment, and the impact of price, but it provides no information beyond fixed configurations of features in a product. Adaptive design only presents relevant alternatives in a conjoint exercise, it generally does not need any prohibitions or conditions. The conjoint part of the test is easier and more straightforward. But if we test innovations, the risk could be in an adaptive CBC's limited ability to model switching from one configuration to another. One more option could be a partial profile design, but in general it has a slightly different set of applications. A partial profile design is mostly used to deal with a large number of attributes and levels, and we are not addressing it in this paper.

CASE STUDY

A case study was executed to understand the impact of a CBC design on the model, results, and conclusions. A simplified version of a real-life project was used as an outline for the case study. It does not exactly recreate the design or the results of this real-life study. The following elements of the original project were used in the case study: the general layout, some attributes, individual utilities for these attributes and the None, and the question of bringing a digital faucet (see Figure 1) from commercial spaces into homes. Bathroom faucets is a competitive category, and the client was considering an interesting innovation. What if the faucets people are familiar with from airports, offices, and restaurants are sold to general consumers? What features and price would optimize revenue in this case? How would the new faucet cannibalize from competitors and from the client's current assortment? To answer these questions, a CBC study was executed. Respondents were screened for being open to buying a digital faucet. The primary attribute was the Water Activation Type (Sensor Only, Handle and Sensor, and Handle Only). The study utilized a custom design with a relatively large number of levels and attributes. Questions about the faucet style and color were asked outside of the CBC and used as covariates in the estimation.

Figure 1: A digital faucet with temperature regulation and display.

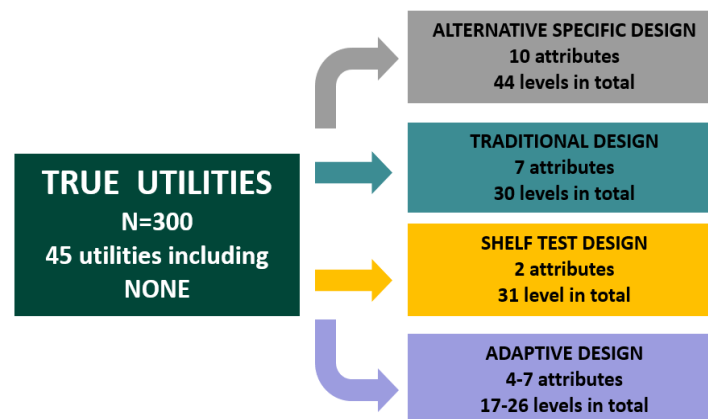


What would be our considerations in choosing the best design for the case study? Different approaches are possible. An alternative-specific CBC seems to be the most natural to test a new digital faucet with different features against standard faucets. Traditional design would be the most familiar for many researchers and could accommodate a complex logic like multiple primary attributes, different conditions, and prohibitions. A shelf test will ensure focus on the impact of price on choices and an adaptive CBC will help verify only the most relevant options for each respondent in a choice exercise.

In the case study, all four models were tested against each other side by side (see Figure 2). To do this, we started with a set of utilities derived from the original real-life study and declared them the “TRUE utilities” reflecting the real preferences of 300 “artificial respondents.” Then, similarly to any other CBC study, designs were generated for all 4 models—alternative-specific, traditional, shelf test, and adaptive CBC—using Lighthouse Studio 9.14.2. For simplicity, the same macro parameters were selected for each of the 4 designs: 50 versions, 14 tasks, 6 alternatives per task, traditional None option. The designs were realistic, they were tested in Lighthouse for balance and statistical efficiency. All 4 designs met the requirements and standards and could be considered satisfactory for a conjoint study.

The next step was to generate choices in conjoint exercises with each design based on the same set of TRUE utilities. For the tests, three sets of choices were generated for each of the four models. The noise in choices was generated using a Gumbel error. Most of the reported results were generated with a relatively high but realistic level of noise (called “Moderate noise” in the paper). The amplitude of the Gumbel error was selected to be approximately 1.5 an average error estimate for the TRUE utilities in the original real-life study (Ye et al., 2017). Application of this level of noise results in about 45% of choices different from the ones generated with TRUE utilities and no noise (“TRUE choices”).

Figure 2: Case Study Setup: Four Models



Designs were generated and tested and then the models were estimated using HB in Sawtooth Lighthouse Studio. The case study was completed with simulating scenarios, running sensitivity analysis and performing optimization to demonstrate strengths and weaknesses of each model reflecting consumers’ preferences and translating them into conclusions and recommendations.

CASE STUDY DESIGN LAYOUTS

Before comparing the models and CBC results, details of the designs for all four conjoint approaches have to be considered.

Table 1: Alternative-Specific CBC Layout

1	Brand	Brand 1	Brand 2	Brand 3	Brand 4	Brand 5
---	-------	---------	---------	---------	---------	---------

2	Water Activation Type	Sensor Only	Handle and Sensor	Handle Only		
3	Handle Location	No	Top Handle	Side Handle	Handle on Base	
4	Sensor Type	Proximity on Front	Proximity Under	Wave on Side	Wave on Top	Touch Sensor
5	Length of Sensor Activation	Deactivate with Sensor	Deactivated After 30 Sec	Deactivated After 1 Min	Deactivate with Handle	
6	Ability to Adjust Water Flow	No	Yes, With Handle	No, Always Full Flow		
7	LED Light to Provide Feedback on Water Temp	Yes	No	No		

Primary Attribute

8 Conditional Attributes

8	Price (Handle Only)	\$75	\$85	\$95	\$105	\$115	\$125	\$135
9	Price (Sensor Only)	\$115	\$125	\$150	\$175	\$200	\$225	\$250
10	Price (Sensor And Handle)	\$200	\$225	\$250	\$275	\$300	\$325	\$350

With the alternative-specific design, 10 attributes and 44 levels were evaluated (see Table 1). In the alternative-specific CBC and in the other three CBC variations, brand is a standard attribute presented on five levels. Water Activation Type on three levels—Sensor Only, Handle and Sensor, and Handle only—is a primary attribute. All other attributes depend on the Water Activation Type. Obviously, Handle Location is only applicable if the faucet has a handle, Sensor Type is only relevant for digital faucets with sensors, the same is true for the Length of Sensor Activation and the LED Light. For the simplified layout in the case study, the handle is always used to adjust temperature, and might be used to adjust water flow (Ability to Adjust Water Flow attribute) in configurations where the handle is present. Like all the other conditional attributes, the Price depends on the Water Activation Type. The Price attribute was not conditioned on Brand, since all the brands in the test were leaders of the category and were offering faucets in approximately the same price range.

For a traditional CBC tested in the simulation, the attribute structure was simplified. The primary attribute (Water Activation Type) was eliminated, and its levels were moved to the Handle Location and Sensor Type attributes (see Table 2). A new level—No Handle—was added to the levels of Handle Type, and No Sensor level is added to the Sensor Type attribute. A traditional CBC design like this would be impossible without prohibitions. To minimize the effect of prohibitions on the design, only completely impossible combinations were excluded, like ability to adjust the water with a handle if a faucet does not have a handle. LED light was allowed in all configurations to avoid extra prohibitions. With this model, we are testing conditional pricing. Conditional pricing was utilized with the traditional CBC. The Price attribute had only seven levels in the design and the correct price was piped into alternatives based on the levels of the Handle Location and Sensor Type attributes.

Table 2: Traditional CBC Layout

1	Brand	Brand 1	Brand 2	Brand 3	Brand 4	Brand 5
2	Handle Location	Top Handle				
		Side Handle				
		Handle on Base				
		No Handle				
3	Sensor Type	Proximity on Front				
		Proximity Under				
		Wave on Side				
		Wave on Top				
		Touch Sensor				
		No Sensor	Prohibit With No Handle			
4	Length of Sensor Activation	Deactivate with Sensor				
		Deactivated After 30 Sec	Prohibit With No Sensor			
		Deactivated After 1 Min	Prohibit With No Sensor			
		No Sensor Activation	Prohibit With No Handle			
5	Ability to Adjust Water Flow	Yes, with handle	Prohibit With No Handle			
		No, always full flow				
6	LED Light to Provide Feedback on Water Temperature	Yes				
		No				

}

Prohibitions

7	PRICE	Price 1	Price 2	Price 3	Price 4	Price 5	Price 6	Price 7
	Price (Handle Only)	\$75.00	\$85.00	\$95.00	\$105.00	\$115.00	\$125.00	\$135.00
	Price (Sensor Only)	\$115.00	\$125.00	\$150.00	\$175.00	\$200.00	\$225.00	\$250.00
	Price (Handle and Sensor)	\$200.00	\$225.00	\$250.00	\$275.00	\$300.00	\$325.00	\$350.00

← Conditional Price

In the shelf test, 15 faucet configurations were evaluated (see Table 3). The configurations were chosen to represent the space of alternatives evaluated in the case study: for each brand, 3 faucets were included, one with every water activation type. All other parameters varied in the 15 configurations. The 15 faucets included in the shelf test were also used in the base case scenario to compare results for all 4 models. Possible prices for all 15 configurations were unfolded into a price grid (see Table 4). In the shelf test, the versions and tasks were designed and balanced for all 16 unique price points across all faucet configurations.

Table 3: Shelf Test: Faucet Configurations

1	Faucet	Faucet 1	Brand 1	Sensor Only	Proximity on Front	Deactivate With Sensor	LED Light		
		Faucet 2	Brand 1	Handle and Sensor	Side of Faucet	Proximity Under	Deactivates After 30 Seconds	Ability to Adjust Water Flow	
		Faucet 3	Brand 1	Handle Only	Top of Faucet	Deactivate With Sensor			
		Faucet 4	Brand 2	Sensor Only	Proximity Under	Deactivates After 1 Minute	LED Light		
		Faucet 5	Brand 2	Handle and Sensor	Top of Faucet	Wave on Side	Deactivates After 30 Seconds	Ability to Adjust Water Flow	LED Light
		Faucet 6	Brand 2	Handle Only	On the Base	Deactivate With Sensor			
		Faucet 7	Brand 3	Sensor Only	Wave on Side	Deactivates After 1 Minute	LED Light		
		Faucet 8	Brand 3	Handle and Sensor	On the Base	Wave on Top	Deactivate With Sensor	Ability to Adjust Water Flow	
		Faucet 9	Brand 3	Handle Only	Side of Faucet	Deactivate With Sensor			
		Faucet 10	Brand 4	Sensor Only	Proximity on Front	Deactivates After 30 Seconds	LED Light		
		Faucet 11	Brand 4	Handle and Sensor	Top of Faucet	Wave on Top	Deactivates After 30 Seconds		
		Faucet 12	Brand 4	Handle Only	On the Base	Deactivate With Sensor			
		Faucet 13	Brand 5	Sensor Only	Touch Sensor	Deactivates After 1 Minute	LED Light		
		Faucet 14	Brand 5	Handle and Sensor	Side of Faucet	Touch Sensor	Deactivates After 1 Minute	Ability to Adjust Water Flow	LED Light
		Faucet 15	Brand 5	Handle Only	Side of Faucet	Deactivate With Sensor	Ability to Adjust Water Flow		

Table 4: Shelf Test: Prices

2	PRICE	\$75	\$85	\$95	\$105	\$115	\$125	\$135	\$150	\$175	\$200	\$225	\$250	\$275	\$300	\$325	\$350
	Price (Handle Only)	\$75	\$85	\$95	\$105	\$115	\$125	\$135									
	Price (Sensor Only)					\$115	\$125		\$150	\$175	\$200	\$225	\$250				
	Price (Sensor and Handle)										\$200	\$225	\$250	\$275	\$300	\$325	\$350

For the case study a simple variation of an adaptive CBC was selected since it is more comparable with the three other designs tested in the study. The simplified adaptive CBC was designed to better demonstrate differences and similarities with the other three models compared to a full-featured ACBC. For the simplified adaptive CBC, three mini designs were created, one for each water activation type (see Table 5); each design has seven tasks with six alternatives per task. Based on the TRUE utilities (with noise or no noise), two out of three water activation types preferred by each respondent were selected and two versions of mini designs for the best water activation types were used to generate choices in the adaptive CBC. For the estimation, all designs and choices were stacked to generate a set of utilities for all levels in the three designs for the Adaptive CBC.

Table 5: Adaptive CBC: Three Mini Designs

1	Brand	Brand 1	Brand 2	Brand 3	Brand 4	Brand 5
---	-------	---------	---------	---------	---------	---------

Sensor Only			Sensor and Handle			Handle Only		
2	Sensor Type	Proximity on Front Proximity Under Wave on Side Wave on Top Touch Sensor	2	Handle Location	Top of Faucet Side of Faucet On the Base	2	Handle Location	Top of Faucet Side of Faucet On the Base
3	Length of Sensor Activation	Deactivate With Sensor Deactivates After 30 Seconds Deactivates After 1 Minute	3	Sensor Type	Proximity on Front Proximity Under Wave On Side Wave on Top Touch Sensor	3	Ability to Adjust Temperature	Yes, With Handle No, Always Full Flow
4	LED Light	Yes No	4	Length of Sensor Activation	Deactivate With Sensor Deactivates After 30 Seconds Deactivates After 1 Minute			
			5	Ability to Adjust Water Temperature	Yes, With Handle No, Always Full Flow			
			6	LED Light	Yes No			

5	Price (Handle Only)	\$75	\$85	\$95	\$105	\$115	\$125	\$135
7	Price (Sensor Only)	\$115	\$125	\$150	\$175	\$200	\$225	\$250
4	Price (Sensor and Handle)	\$200	\$225	\$250	\$275	\$300	\$325	\$350

CASE STUDY RESULTS

Choices generated for each of the 300 “artificial respondents” in all 14 tasks with the four tested designs (alternative-specific, traditional, shelf test, and adaptive) were used to estimate models, simulate shares of preference, and run sensitivity analysis for faucet features and price.

To evaluate the models’ fit overall, the hit rate was estimated (see Table 6). For the hit rate test, three levels of noise were used. First, the four models were estimated with no noise in choices (TRUE choices). Clearly, no noise is not realistic since it does not include error levels in the data. As discussed above, Moderate noise is the most realistic since it mimics the average individual error estimate in the real-life model the case study is based on. A higher level of noise (“High Noise” in the table) was mostly used for a stress test, to see the point where the models break. High noise corresponded to about 65% of choices different from the TRUE choices. It means that respondents didn’t see a lot of difference between the alternatives and often made random choices.

Since TRUE choices were known for every “artificial respondent,” the hit rate was calculated in all 14 tasks and no fixed tasks were used. The hit rates with TRUE choices and with choices

with noise are somewhat similar across the models and relatively high even in presence of realistic Moderate noise. The alternative-specific model has the best hit rate at various levels of noise. And the Shelf Test even shows some denoising properties. Hit Rate compared to the TRUE choices is the best for the shelf test.

Table 6: Hit Rates for the Four Models

	Alternative Specific CBC			Traditional CBC with Prohibitions		
	No Noise	Moderate Noise	High Noise	No Noise	Moderate Noise	High Noise
Hit Rate	99.6%	84.0%	66.9%	98.8%	70.8%	61.3%
Hit Rate with TRUE Choices	99.6%	68.4%	44.2%	98.8%	59.5%	42.9%

	Shelf Test			Adaptive CBC		
	No Noise	Moderate Noise	High Noise	No Noise	Moderate Noise	High Noise
Hit Rate	99.50%	73.60%	54.90%	94.90%	72.50%	49.30%
Hit Rate with TRUE Choices	99.50%	73.90%	51.10%	94.90%	61.00%	31.90%

To further investigate differences between the models, attribute importance scores were estimated for each model. The importance scores indicate the impact of every attribute on choice within a model. In the case study the TRUE utilities are known and are the same for all four models, the attributes and the number of levels is similar across the models, so there is an opportunity to compare importance scores calculated based on the estimated utilities with the importance scores calculated based on the TRUE utilities (“TRUE importance”). The results of this comparison are summarized in Table 7 and Table 8. After the estimation, the alternative-specific CBC significantly exaggerates the importance of the primary attribute—Water Activation Type. The TRUE importance of the primary attribute is high, but after the estimation with no noise it becomes even higher, and the Price attribute importance score is suppressed the most among all the attributes. Even though the primary attribute is not directly presented in the traditional CBC, the estimation has the same effect on the importance scores as with the alternative-specific model. The attributes related to the water activation type—Handle Location and Length of Activation—get more importance and the Price importance is dramatically suppressed after the estimation. To summarize, both the alternative-specific design and the traditional design drive attention to the primary attribute or to the attribute/s defining prohibitions and by design the Price (depending on the primary attribute/s) becomes less important.

The noise in simulated choices makes the difference between attribute level utilities smaller, driving the average estimated utilities closer to zero. There is a chance for a moderate noise to smooth down the importance score based on estimated utilities for the primary attribute and make it closer to the TRUE importance score. The effect is observed with the alternative-specific model which has a well-defined primary attribute and is more robust in the presence of noise. For the traditional model with Moderate noise, the importance imbalance becomes worse, especially for Price.

Table 7: Importance Scores Comparison: Alternative-Specific CBC and Traditional CBC**Alternative-Specific CBC**

	Brand	Water Activation Type	Handle Location	Sensor Type	Length of Activation	Ability to Adjust Flow	LED Light	Price
TRUE Utilities	11%	29%	8%	11%	10%	7%	5%	18%
Estimated, No Noise	9%	38%	7%	10%	11%	8%	4%	12%
Estimated, Moderate Noise	10%	34%	8%	11%	10%	7%	5%	15%
Estimated, High Noise	14%	25%	10%	15%	11%	7%	5%	13%

Traditional CBC

	Brand	Handle Location	Sensor Type	Length of Activation	Ability to Adjust Flow	LED Light	Price
TRUE Utilities	13%	14%	26%	12%	9%	6%	21%
Estimated, No Noise	11%	18%	35%	12%	8%	5%	10%
Estimated, Moderate Noise	13%	20%	31%	13%	9%	6%	8%
Estimated, High Noise	17%	20%	28%	12%	9%	6%	7%

The issue with the imbalance of the attribute importance described above reflects the transformation of the high-level distribution after the CBC estimation: independently of the model, the attribute level utilities' averages after the estimation stay almost the same compared to the TRUE utilities' averages, but the variance of level utilities distribution becomes significantly higher for the primary attribute or for the attribute/s defining other attributes.

The importance scores for the shelf test and for the adaptive CBC are compared in Table 8. The shelf test is the best in preserving the ratio between the importance scores of the two attributes it evaluates. The shelf test only considers 15 fixed configurations, so the Price becomes a clear differentiator and its impact on choices is estimated accurately. In the adaptive CBC tested in the case study, the importance of the Price attribute is significantly understated after the estimation, since by design respondents were not given an opportunity to switch from one water activation type to another based on price in the conjoint exercise.

Table 8: Importance Scores Comparison: Shelf Test and Adaptive CBC**Shelf Test**

	Faucet	Price
TRUE Utilities	76%	24%
Estimated, No Noise	80%	20%
Estimated, Moderate Noise	78%	22%
Estimated, High Noise	74%	26%

Adaptive CBC

	Brand	Water Activation Type	Handle Location	Sensor Type	Length of Activation	Ability to Adjust Flow	LED Light	Price
TRUE Utilities	11%	29%	8%	11%	10%	7%	5%	18%
Estimated, No Noise	12%	28%	9%	13%	14%	10%	5%	9%
Estimated, Moderate Noise	14%	22%	10%	15%	14%	9%	7%	9%
Estimated, High Noise	19%	10%	12%	20%	14%	9%	8%	9%

ANALYSIS: BASE CASE SCENARIO AND SHARES OF PREFERENCE

To further investigate the impact of modeling in the case study, the same base case scenario was simulated for the 4 models and the shares of preference in this scenario were estimated. For the simulations, an extensive scenario with all 15 faucets from the shelf test was selected (see Table 3); different price levels were selected for different faucets by different brands (see Table 9). A None option was included in the base scenario.

Table 9: Prices in Base Case Scenario

	Price (Handle Only)							Price (Sensor Only)							Price (Handle and Sensor)						
	\$75	\$85	\$95	\$105	\$115	\$125	\$135	\$115	\$125	\$150	\$175	\$200	\$225	\$250	\$200	\$225	\$250	\$275	\$300	\$325	\$350
Faucet 1									+								+				
Faucet 2																					
Faucet 3		+																			
Faucet 4										+											
Faucet 5																		+			
Faucet 6			+																		
Faucet 7											+										
Faucet 8																			+		
Faucet 9				+																	
Faucet 10												+									
Faucet 11																				+	
Faucet 12					+																
Faucet 13													+								
Faucet 14																					+
Faucet 15						+															

The share of preference for the four models tested in the case study with utilities estimated with Moderate noise is summarized in Table 10. The shares based on the models with no noise (“TRUE shares”) are summarized above for reference. **Even with the extensive base case scenario and in presence of noise, all four models were mostly aligned in estimating the shares of preference and identifying the best and the worst configurations.**

**Table 10: Share of Preference in Base Case Scenario:
TRUE Utilities and Estimated Utilities with Moderate Noise**

Share of Preference TRUE Utilities	Faucet 1	Faucet 2	Faucet 3	Faucet 4	Faucet 5	Faucet 6	Faucet 7	Faucet 8	Faucet 9	Faucet 10	Faucet 11	Faucet 12	Faucet 13	Faucet 14	Faucet 15	None
Alternative Specific CBC	9.8%	7.0%	5.4%	4.6%	11.8%	5.6%	3.6%	9.2%	4.5%	4.3%	5.3%	4.9%	3.4%	6.2%	11.1%	3.2%
Traditional CBC	9.8%	6.9%	5.5%	4.3%	9.9%	5.6%	3.4%	8.4%	4.6%	4.1%	5.1%	5.0%	2.7%	5.6%	11.0%	8.3%
Shelf Test	10.0%	7.1%	5.4%	4.7%	11.8%	5.6%	3.6%	9.2%	4.5%	4.4%	5.3%	4.9%	3.5%	6.3%	11.2%	2.5%
Adaptive CBC	9.8%	7.0%	5.4%	4.6%	11.8%	5.6%	3.6%	9.2%	4.5%	4.3%	5.3%	4.9%	3.4%	6.2%	11.1%	3.2%

Share of Preference Estimation, Moderate Noise	Faucet 1	Faucet 2	Faucet 3	Faucet 4	Faucet 5	Faucet 6	Faucet 7	Faucet 8	Faucet 9	Faucet 10	Faucet 11	Faucet 12	Faucet 13	Faucet 14	Faucet 15	None
Alternative Specific CBC	12.4%	6.3%	5.8%	3.4%	9.7%	4.8%	3.0%	10.7%	3.6%	4.9%	5.5%	4.0%	3.4%	8.4%	12.7%	1.3%
Traditional CBC	11.3%	6.7%	10.4%	3.9%	8.7%	6.0%	4.7%	8.4%	4.6%	4.7%	5.4%	4.3%	3.3%	5.9%	5.8%	6.1%
Shelf Test	9.4%	6.8%	5.8%	4.9%	13.3%	6.6%	3.6%	9.3%	3.8%	3.4%	4.8%	6.1%	3.1%	6.3%	10.7%	2.1%
Adaptive CBC	9.9%	6.1%	5.3%	4.8%	11.2%	5.6%	5.2%	10.4%	4.8%	4.1%	4.3%	4.9%	4.1%	6.5%	10.2%	2.7%

The shelf test is the best in recreating the TRUE shares after the estimation with Moderate noise. It is closely followed by adaptive CBC. The error in shares estimation is still acceptable but significantly higher for alternative-specific CBC (see Table 11).

**Table 11: Share of Preference Error:
Maximal and Average Absolute Difference Between the True Share
and the Share Based on Utilities Estimated with Moderate Noise**

	MAX Abs Diff	AVERAGE Abs Diff
Alternative Specific CBC	2.6%	1.1%
Traditional CBC	5.2%	1.2%
Shelf Test	1.5%	0.5%
Adaptive CBC	1.6%	0.5%

The alternative-specific CBC focuses on features, especially on the primary attribute, but unlike the shelf test and the adaptive CBC it collects less data and allows fewer pairwise comparisons for the particular faucet configurations in the choice exercise. Traditional CBC is comparable but slightly worse than the alternative-specific CBC in share of preference estimation with the Moderate noise. Even if simulated with TRUE utilities, the share of None is

significantly higher for the traditional CBC. That arises because the lack of conditional relationships between the attributes in the traditional CBC generates more unacceptable alternative sets. It might explain some discrepancy in shares after the estimation for the traditional CBC. Researchers should keep in mind that a share of None could be overstated in estimation with a traditional CBC design if complex relationships between attributes are present. It is important to take it into account when a proper CBC layout and design are selected.

SENSITIVITY ANALYSIS FOR FEATURE OPTIMIZATION

The outcome of sensitivity analysis was compared in the case study for the three models: alternative-specific CBC, traditional CBC, and adaptive CBC. The shelf test is excluded from this comparison since it does not estimate utilities separately for each feature. For feature optimization in the case study the sensitivity analysis was performed in a simplified scenario with only one faucet against None. Two sensitivity tests were executed (see Table 12). In the first one, the levels of the primary attribute—Water Activation Type—were varied and the share of preference was recorded for the base case configuration. In this case, prohibited levels are excluded from the configuration based on the Water Activation Type and the Price is set to the lowest level in the corresponding price interval. For the second sensitivity test, the levels of the Sensor Type attribute were varied for the same Handle and Sensor faucet.

Table 12: Base Case Scenarios for Feature Optimization

Primary Attribute:
Water Activation Type

	Brand	Water Activation Type			Handle	Sensor Type					Deactivation	Ability to Adjust Flow	LED Light	Price		
	Brand 1	Sensor Only	Handle and Sensor	Handle Only	Side Handle	Proximity on Front	Proximity Under	Wave on Side	Wave on Top	Touch Sensor	After 30 Sec	Yes	No	\$75	\$115	\$200
Scenario 1	+		+		+		+				+	+	+			+
Scenario 2	+	+					+				+	+	+		+	
Scenario 3	+			+	+									+		

Conditional Attribute: Sensor Type

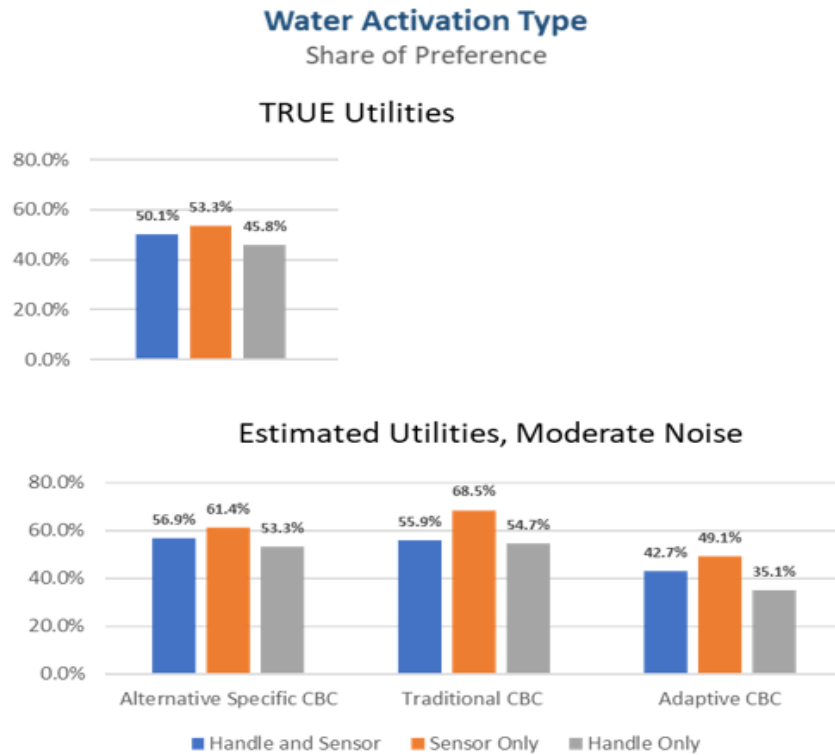
	Brand	Water Activation Type			Handle	Sensor Type					Deactivation	Ability to Adjust Flow	LED Light	Price		
	Brand 1	Sensor Only	Handle and Sensor	Handle Only	Side Handle	Proximity on Front	Proximity Under	Wave on Side	Wave on Top	Touch Sensor	After 30 Sec	Yes	No	\$75	\$115	\$200
Scenario 1	+		+		+	+					+	+	+			+
Scenario 2	+		+		+		+				+	+	+			+
Scenario 3	+		+		+			+			+	+	+			+
Scenario 4	+		+		+				+		+	+	+			+
Scenario 5	+		+		+					+	+	+	+			+

The sensitivities based on estimated utilities with Moderate noise for different models are presented in Figure 3. The TRUE sensitivity to the Water Activation Type is presented on the top for a reference. The sensitivity analysis based on the estimated utilities in the framework of the alternative-specific CBC was the closest to the TRUE sensitivity; the traditional CBC overestimated the share of the strongest Water Activation Type level—Sensor Only, and the adaptive CBC suppressed the share of the primary attribute by design, but still recreated the relative strength of the three levels of the Water Activation Type.

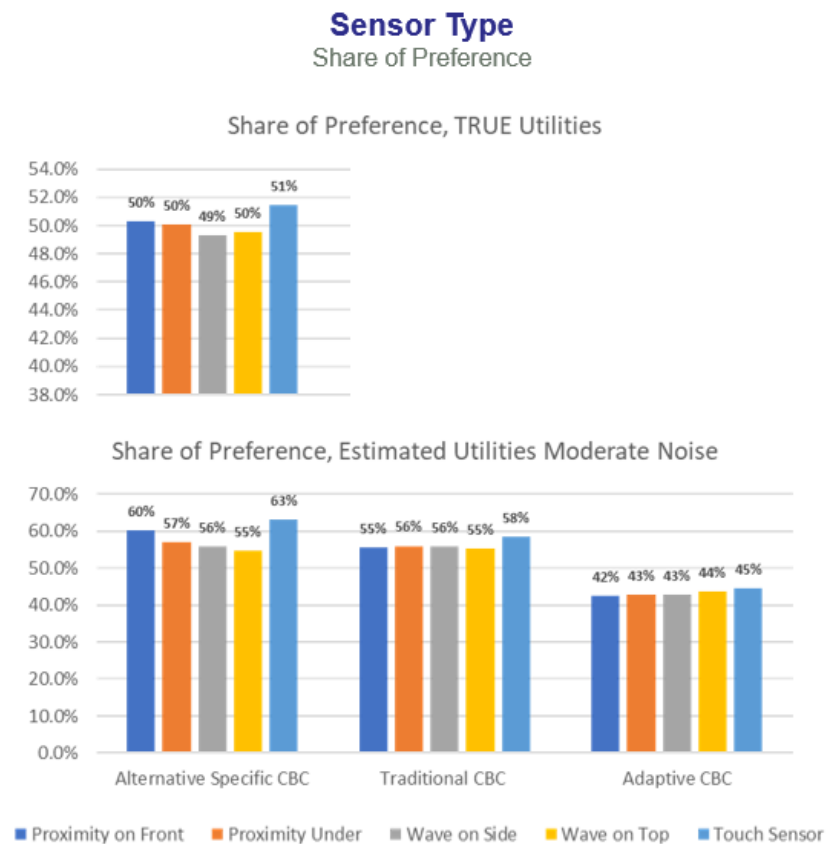
The results of the sensitivity analysis for one of the conditional attributes—Sensor Type—are presented in Figure 4. Again, after the estimation with Moderate noise, the alternative-specific CBC is the most accurate in recreating the TRUE sensitivity across the levels of the conditional attribute, the traditional CBC correctly indicates the strongest level, and the adaptive CBC

suppresses sensitivity across all five levels. Overall, the alternative-specific CBC demonstrates a better fit for feature optimization than the other two approaches.

**Figure 3: Sensitivity Analysis:
Share of Preference in the Base Case Scenario for Levels of Water Activation Type**



**Figure 4: Sensitivity Analysis:
Share of Preference in the Base Case Scenario for Levels of Sensor Type**

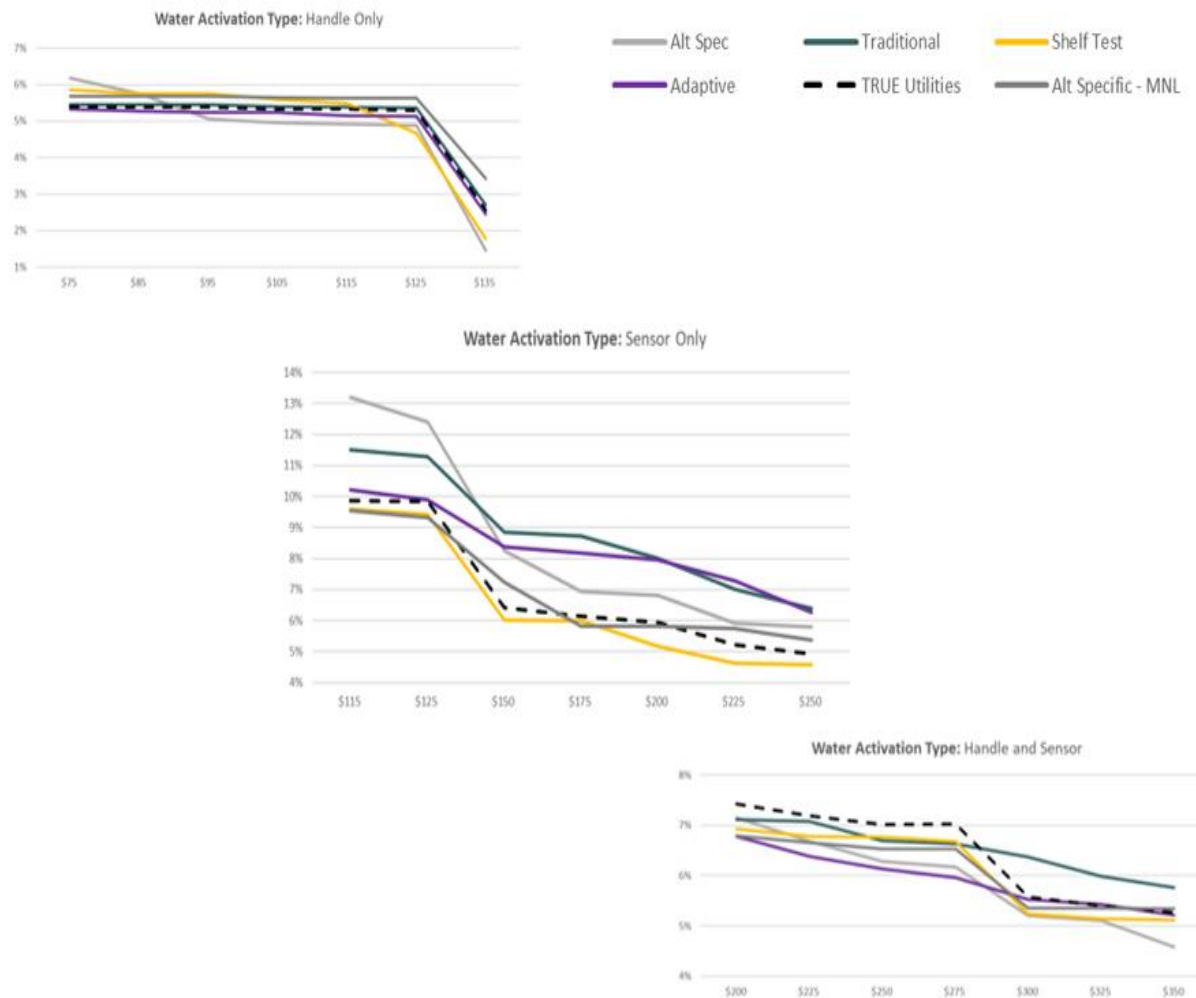


PRICE SENSITIVITY ANALYSIS

Price sensitivity analysis was performed for all four models in the full base case scenario with 15 faucet configurations (see Table 3 for the configurations and Table 9 for the prices in the base case scenario). Price sensitivity based on estimated utilities for each model was compared with the TRUE sensitivity. Figure 4 displays the price sensitivity curves for the three faucets in the base case scenario—Faucet 1, Faucet 2, and Faucet 3. These faucets are all offered by Brand 1, and the main difference between the faucets is the primary attribute—Water Activation Type. In all models tested in the case study the price is conditional on the Water Activation Type. Handle Only corresponds to the lowest price interval, Sensor Only is in the middle price interval, and Handle and Sensor is in the highest price interval.

With Moderate noise, all four models are able to mimic the TRUE price sensitivity curve for all three Water Activation Types. The shelf test sensitivity curve aligns with the TRUE sensitivity very closely. Due to its simplicity, the shelf test focuses the most on the trade-offs based on the price. The sensitivity curves built with the traditional CBC with conditional pricing and with the adaptive CBC are too flat in the highest price interval.

**Figure 4: Price Sensitivity Curves:
Share of Preference in the Base Case Scenario for Three Faucets
with Different Water Activation Type**



The sensitivity curve built with the alternative-specific CBC deviates the most from the TRUE price sensitivity for all three water activation types. It is especially noticeable in the middle price interval—with the most popular Sensor Only water activation type. In the middle price interval, the share of preference is significantly overstated on the left side of the interval (for lower prices) and understated on the right side of the interval (for higher prices), making the sensitivity look higher compared to the TRUE sensitivity in this price interval. The same effect is demonstrated with the alternative-specific design in the lowest and highest price intervals but is less pronounced.

Various approaches were tried to improve the price sensitivity analysis outcomes for the alternative-specific design. These included using other design versions, varying design parameters (Chrzan and Orme, 2000), varying the prior density, using linear price instead of part-worths (Orme, 2007), and using the first-choice model instead of the probabilistic one. With all the modifications, the overstated sensitivity issue was still visible in the framework of the alternative-specific model. The issue seems to be related to the HB estimation building a top-level model in the alternative-specific context. The issue is not present if a Multinomial Logit

estimation (MNL) is used for the alternative-specific model (see Figure 40). Researchers should pay attention to this issue with the HB for alternative-specific CBC and might want to use an MNL approximation to calibrate the sensitivity curves in this case.

CONCLUSIONS

All four models tested in the case study—alternative-specific CBC, traditional CBC, shelf test, and adaptive CBC—showed remarkable accuracy of estimation and stability in presence of noise. All of them could be applied to solve particular business problems, including optimization in tech and durable goods categories.

Moreover, the models were sophisticated and detailed enough to underline differences in respondents' reactions to choice exercises based on different CBC designs. The alternative-specific CBC and the traditional CBC draw attention to particular features, especially the primary attribute or the attribute/s defining conditions and prohibitions in a design. The shelf test focuses on a limited number of fixed configurations and the configuration/price trade-offs. And the adaptive CBC studies configurations and prices that are the most relevant to each respondent but has limited ability to capture switching behavior.

These properties of each CBC design dictate pluses and minuses of each of the methods for analysis in complex categories. Adaptive CBC is flexible and relevant but might have limited applications in testing innovations. Traditional CBC is standard and familiar to most researchers; it can recreate some relationships between attribute levels with conditions and prohibitions, but it might show weaknesses in estimating importance, sensitivity, and a None level. Shelf test is very accurate in estimating shares of preference and price sensitivities, but it is only limited to testing fixed configurations and cannot be used for feature optimization.

Alternative-specific CBC could be the best fitting conjoint type for many business problems in complex categories such as tech and durable goods. It naturally recreates the situation where levels of a primary attribute define other attributes in a product or offering and is suitable for testing and optimizing innovations. If a product or offering is described with a hierarchy or primary and conditional attributes, an alternative-specific CBC would be the best approach for feature optimization. A researcher has to be careful with importance score estimation and interpretation and with price sensitivity analysis in the framework of an alternative-specific CBC.

As shown in the case study, the sensitivity to price might be overstated especially at the ends of the tested price interval if the price is conditioned on the primary attribute in an alternative-specific CBC. Additional research is needed to see if the problem can be resolved with design, regularization, or calibration.



Faina Shmulyian



Tyler Dugan

REFERENCES

- Chrzan, Keith and Bryan Orme. *An Overview and Comparison of Design Strategies for Choice-Based Conjoint Analysis*. Sawtooth Software Research Paper Series. (Sawtooth Software, Inc.: 2000).
- Liu, YiChun Miriam, Jeff Brazell and Greg Allenby. *An Integrative Model for Complex Products*. Sawtooth Software Conference Proceedings (Sawtooth Software, Inc.: 2021).
- Orme, Bryan. *Three Ways to Treat Overall Price in Conjoint Analysis*. Sawtooth Software Research Paper Series. (Sawtooth Software, Inc.: 2007).
- Ye, Xin, Venu M. Garikapati, Daehyun You and Ram M. Pendyala. A Practical Method to Test the Validity of the Standard Gumbel Distribution in Logit-Based Multinomial Choice Models of Travel Behavior. *Transportation Research Part B* 000 (2017) 1–20.

CLUSTERING OPEN-ENDED QUESTIONS: THE ALGORITHM TO AUTOMATICALLY QUANTIFY SPEECH

FEDERICO ADROGUE
KNACK RESEARCH

1. INTRODUCTION

Analyzing open-ended questions in the most efficient, fast and automated way possible is one of the great challenges we still face in quantitative research. The purpose of this paper is to propose an algorithm that performs all of these tasks and delivers results with valuable insights in a matter of minutes.

To achieve these results, we will combine descriptive statistics techniques with machine learning. With the fusion between these methodologies, we will quantify speech; first converting words into numbers, then segmenting the respondents, and finally generating automatic phrases that explain what they are talking about in each segment.

To illustrate this proposal, we will present the results of a study that KNACK did for UNICEF in 2020. In this study we asked young people throughout Argentina to describe in detail what skills they had developed during the quarantine period of the COVID pandemic.

The results were surprisingly clear and insightful. This demonstrates the potential this tool can have and the considerable distance we still have to go in this area of research.

2. CLARIFICATION

Encoding open-ended questions can be one of the most tedious and time-consuming jobs in quantitative research. Have you ever asked yourself whether there is a more efficient way to understand respondents' natural speech? The goal of this new methodology is to attempt to answer that question and save long hours of manual work, getting more accurate results in a matter of minutes.

This algorithm is not simply an automated version of encoding open-ended questions, but rather a different type of analysis that seeks to replace that methodology. While encoding open-ended questions has its own benefits, the exercise can be time-consuming and resource intensive to analyze. This model, on the other hand, offers a streamlined and efficient way to identify key themes and patterns in respondents' answers.

It's worth noting that there has been previous exploration in this area within market research. The most popular method is sentiment analysis, which aims to classify responses based on their positive or negative sentiment. However, sentiment analysis has its limitations and does not accomplish the final task that we are seeking; to segment our respondents based on their areas of interest. This is where this model comes in, providing a more comprehensive approach to understanding respondents' preferences and behaviors.

All of the algorithm programming is created in Python, one of the most powerful languages for this type of analysis. It should be noted that the methodology outlined below is a description of the algorithm's steps. None of these steps require any manual intervention, you only need to run the code.

The algorithm is designed under the Unsupervised Learning methodology, one of the branches within Machine Learning. This means that we are performing an exploratory analysis on the data; we want to segment our cases into clusters based on response patterns that the algorithm perceives. We don't know these patterns beforehand, but the algorithm will help us decipher them so that we can make our own conclusions.

3. METHODOLOGY

In 2020, the pandemic significantly altered the routines and behaviors of younger generations. UNICEF, in conjunction with KNACK, decided to explore what adolescents in Argentina had learned outside of school during the quarantine period. It is worth saying that KNACK has an extensive working experience with UNICEF as a provider of social research studies, including conducting the Multiple Indicator Cluster Surveys (MICS) in Argentina.

This specific project is a great example of the use of the tool. Our goal is to be able to understand the discourse from a quantitative point of view. By focusing on a methodology based on a quantitative analysis, we can leave aside the qualitative approach, as it would be difficult to implement the latter methodology with a large number of respondents.

Data Collection

The sample frame for the study was drawn from across the entire country of Argentina, with over 3,000 cases collected throughout the nation. However, for the specific question that we are analyzing, responses were gathered from 400 adolescents.

In order to explore the young people's responses in depth, we first asked them to write their answers freely and spontaneously, with no word limit:

What do you think you learned beyond school during this period?

A large, empty rectangular text box with a thin grey border, intended for users to provide their answers to the question above. It occupies a significant portion of the lower half of the page.

All responses were stored in a structured database, as written by the respondents:

ID	Question
1	programas de computadoras como excel, powerpoint
2	a pasar tiempo con mis hermanos, estabamos muchas horas jugando y compartiendo
3	a usar mejor la computadora
4	aprendi a cocinar rico
5	hago las tareas del hogar que antes no
6	a valorar a mi papa
7	estar y jugar con sus hermanos
8	a cocinar carnes al horno
9	tenia que lavar los platos todos los dias y antes no sabia ni como lavarlos, tambien ayudaba en i
10	a lavar los pisos de la casa
11	aprendi a usar todos los programas de videollamada como meet zoom y esos
12	a estar con mis amigos de manera online y virtual
13	a ordenar y lavar y ayudar en las tareas del hogar porque me obligaban
14	a ganar en el fornite, muchas horas en la computadora
15	a realizar las tareas del hogar
16	a hacer las cosas y tareas del hogar pero no me gustaba
17	a disfrutar en rutina con toda mi familia
18	a cocinar platos nuevos
19	aprendi de todo pero ahora uso mejor la computadora que antes, gracias a los profesores que i

Data Preparation

Stopwords

Once we had the final database with all of the answers, we needed to eliminate any words that generated “noise” and did not contribute value to the analysis.

Stopwords are common words in a language that do not contribute semantic meaning and can therefore generate noise in a language statistical analysis. We must remove them before implementing our algorithm, as they have no significant impact on the meaning of the text and can affect the accuracy of the results.

To do this, we used the NLTK library in Python (<https://www.nltk.org/>), which allows us to import a list of stopwords, i.e., any words we do not need when analyzing the information. Here is an example with some of the words that integrate the stopwords list: [“de,” “la,” “que,” “el,” “en,” “y,” “a,” . . .]. In this case, we can see that the list of words is in Spanish, but we could apply the same tool in multiple languages. NLTK includes stopwords in several languages, including English, Spanish, French, German, Italian, Portuguese, and many others.

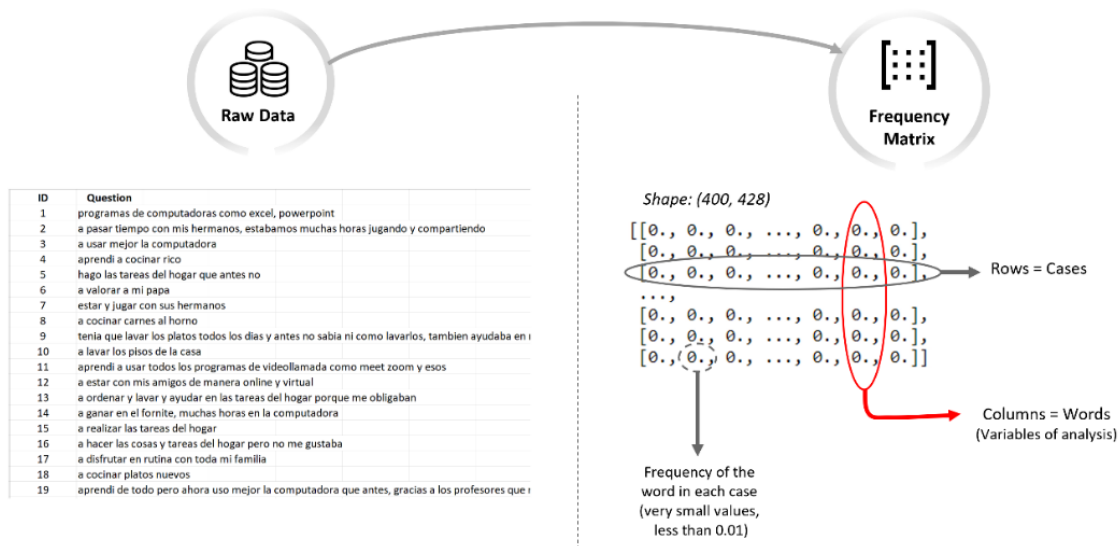
This list automatically goes through each of the respondents’ answers and eliminates the words that do not provide any value.

An example of how a sentence looks after applying stopwords is as follows: “*During quarantine, I learned to appreciate the simple things in life*” and after applying NLTK’s stopwords it would become “*quarantine, learned appreciate simple things life.*”

TFIDF Vectorizer

After removing the stopwords, the next step is to convert the remaining words into numbers to run our statistical analysis.

From the Scikit-Learn library (<https://scikit-learn.org/>), one of the main machine learning tools in Python, we import the TF-IDF or Term Frequency Inverse Document Frequency function. This function allows us to transform our base of open-ended answers into an array, where each value represents the frequency of repetition of each of the words that exist in the database:



As seen in the image, each of the rows represents the respondents' answers, and each of the columns represents a specific word. From the shape of the matrix, we can see that we have 400 rows (one for each respondent) and 428 columns (one for each word that exists in the database).

While 428 features may seem relatively low for an NLP (*Natural Language Processing*) analysis, it is important to note that this can vary greatly depending on the specific research question and dataset at hand. In many cases, NLP analyses may involve working with thousands of unique words, each of which can represent a separate feature or variable. In the case of this study, the research question was highly specific and focused on a narrow set of keywords and phrases. As a result, the number of unique words and features in the dataset was relatively small, which allowed for a more focused and targeted analysis. With a greater number of features and respondents, the segments could be more general and less specific, but we suggest conducting further research to corroborate these hypotheses.

TF-IDF is a way to measure how unique or rare a word is in a particular text or set of texts. It does this by comparing how many times a word appears in one text with how many texts the word appears in total. The more often a word appears in many different texts, the less unique it is considered to be, while the fewer texts it appears in, the more unique it is considered to be.

In the matrix in the previous photo, we can see that the values are less than one; this is because it is a database with a large number of words, where some groups of people talk about very specific topics and the rest about other topics. This marked differentiation will serve us later when forming the clusters.

To learn in depth how the TF-IDF calculations are made, please refer to the following documentation:

https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

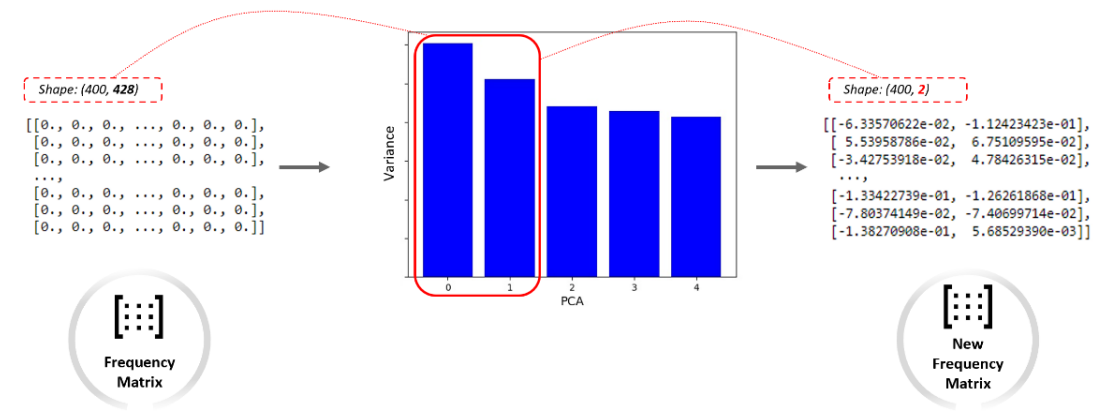
Principal Component Analysis

We have already managed to convert our open-ended response database into a numerical matrix. However, we need to further reduce the dimensions in order to segment our respondents. This step is important to continue reducing the noise in our database and only keep the information that will be useful for conducting the cluster analysis.

For this, we will use the PCA function, which is also in the Scikit-Learn Python library. This function allows us to perform two operations; first to transform the data for “decorrelation” and second, to reduce the dimensions.

The first step transforms our data matrix as follows: the rows still represent individual cases, but the columns now represent the “PCA features.”

In the second step we need to select the PCA features that contain a higher percentage of variance over the data. We will then reduce our entire frequency matrix to the number of chosen dimensions: in this graph the first column represents PC 0 and the second column PC 1:



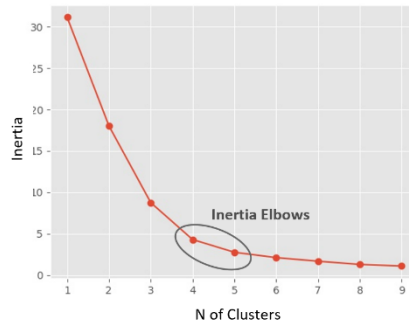
Reducing 428 dimensions to just 2 dimensions will allow us to clean up any “noise” that may still be in our database. Additionally, it will be useful for data visualization and modeling.

It is important to clarify that for the purpose of this paper, only the first two principal components were selected to make it easier to observe the data and visualizations, even though they may not represent the best variance for this particular case. The number of principal components can be increased, and the optimal amount can be selected for each analysis. By increasing the number of principal components, a greater proportion of the variability in the data can be captured, which can lead to more accurate insights.

Choosing the Number of Clusters

So far, we have managed to convert our entire database into a matrix of 400 rows and only 2 columns [shape = (400,2)]. We now have our data ready to segment our respondents using a k-means clustering analysis. However, first we need to select the optimal number of clusters for our sample.

Calculating the inertia of the data will allow us to measure how scattered the data are from each other by measuring the distance between each cluster centroid. A good clustering has tight clusters, so low inertia, but not too many. That is why we choose the point where the inertia starts to decrease, in this case 5 clusters:



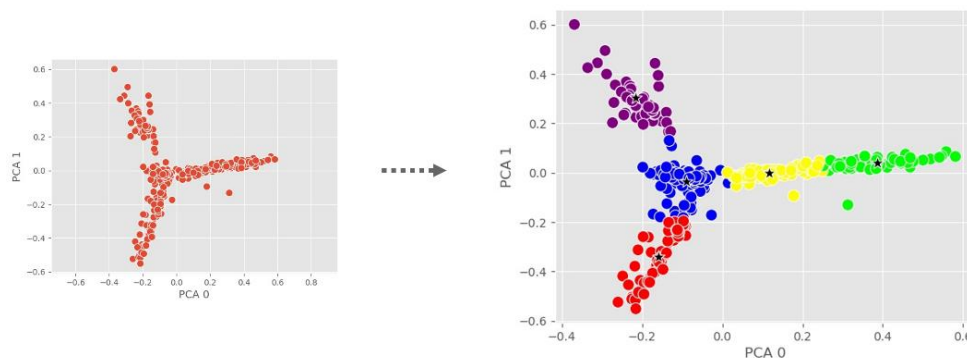
Nevertheless, it is very important once we have reached the final results of our k-means analysis to check if the clusters are adequately representing the information in our data. If we see that there could be a group talking about a topic that is not clearly defined in the current results, we can re-run the analysis including an extra cluster and compare the results. Similarly, if we see that there are two groups talking about very similar topics, we can reduce the number of clusters and compare the results to ensure that we are properly segmenting our respondents.

While we can use the elbow method to determine the optimal quantity of clusters, selecting the cluster count that best represents your sample always requires a touch of art and vision.

Clustering the Data

We now have our dimensionally reduced matrix and the optimal number of clusters for our data (5).

To perform the cluster analysis, we will use the k-means function, which is also from Scikit-Learn. The function allows us to set the number of clusters we want to create and the maximum number of iterations it can have, i.e., how many times it rearranges the centroids to create the most representative segments possible. In this case there are 5 clusters and 2000 iterations:



The visualization of the clusters helps us segment the responses and visualize the most related conversation topics. With this technique, we can identify patterns and group similar responses, allowing us to get a clearer and more precise understanding of the data.

Additionally, the visualization of the clusters also allows us to determine which topics are closest to each other, which can be useful for identifying important trends and patterns in the conversation. However, visualizing the clusters becomes increasingly difficult as the number of dimensions increases.

NMF (Non-Negative Matrix Factorization)

Next, we need to understand what topics are being talked about in each of the segments we created. To accomplish this, we first have to divide our original database with the respondents' wording into five parts (one per cluster).

Once the database is divided into five parts, we will transform the words back into numbers as we did in the first step. The difference will be that, instead of having a single frequency matrix, we will have five different frequency matrices, each one corresponding to a cluster.

It is important to note that we are using the same matrix that we previously generated in the TFIDF step. This matrix had 400 rows and 428 columns. However, instead of being a single matrix, it has been divided into 5 matrices, each with a number of rows equal to the number of cases in each cluster. Each matrix contains the responses of the cases corresponding to each cluster.

In order to understand which words have the greatest impact in each of the five matrices, we will use NMF (Non-Negative Matrix Factorization) analysis. This is a widely used function in machine learning that will allow us to identify what consumers are talking about and which are the topics that stand out the most.

Here you can see the main words per cluster that the NMF analysis gives us in this case:

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
computadora 0.912322	valorar 0.898445	amigos 0.839633	limpiar 0.995546	cocinar 1.071924
aprendi 0.682216	mama 0.717038	videollamadas 0.410001	casa 0.709124	rico 0.820141
programar 0.669633	tiempo 0.714862	virtual 0.385896	lavar 0.656055	platos 0.545436
usar 0.424431	hermanos 0.634114	manera 0.379874	ropa 0.561334	nuevos 0.312303
mejor 0.373199	familia 0.631249	jugar 0.360557	ayudar 0.453367	aprendi 0.272768
celular 0.372674	papa 0.603568	valorar 0.356237	planchar 0.410448	cosas 0.257028
videos 0.310898	mas 0.483596	online 0.346942	tareas 0.362255	carnes 0.213125
cursos 0.286163	disfrutar 0.293771	compu 0.320865	cosas 0.298896	asados 0.211500
uso 0.271958	hace 0.236764	no 0.297598	pisos 0.250710	tortas 0.207071
hacer 0.233899	amigos 0.234469	colegio 0.272318	hogar 0.230795	nuevas 0.192224
sociales 0.233736	relaciones 0.232802	hacer 0.266677	ordenar 0.193285	mas 0.179840
redes 0.233736	amistades 0.159388		hacer 0.180198	
tecnologia 0.169803			platos 0.173941	

The function is programmed to display up to 25 words that are most representative of each cluster. The words that appear at the top are the ones that most characterize each group. If enough information is not obtained with the displayed words, the range can be expanded to obtain a longer list.

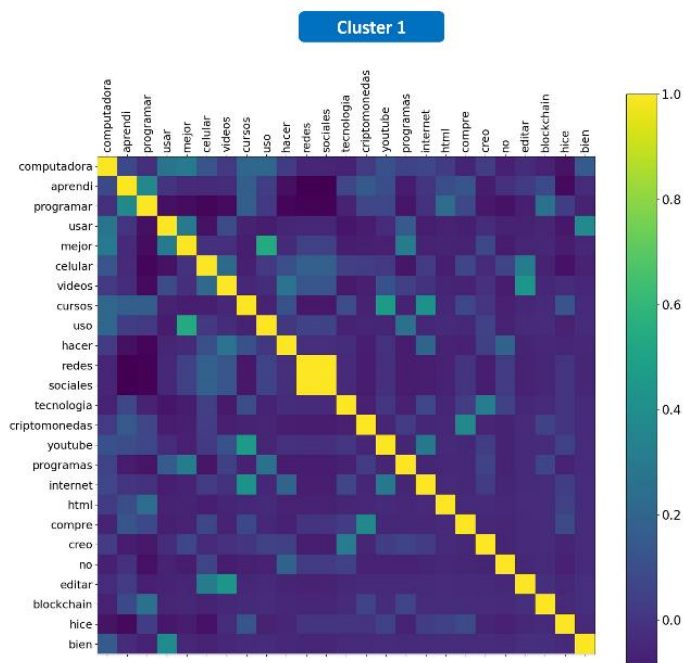
To learn more about how NMF calculations are done, review the following documentation: <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html>

Correlation Matrices

The NMF analysis implemented in the previous step allows us to understand what topics are discussed in each of the clusters, which already gives us a clear understanding of the results.

However, we can still take one more step to gain a better understanding of each cluster. This last step will be to create correlation matrices. In this case, we must create 5 correlation matrices, since each segment will correspond to one, and each matrix will be formed by the top 25 words that best represent each cluster. Keep in mind that we already know the top 25 words per segment since they are the result of the NMF analysis in the previous step.

These 25 words are going to be represented in rows, and then repeated as columns. The intersection between each word calculates their correlation. We represent these matrices as heat matrices, where lighter colors represent a higher correlation. Here is an example of the heat matrix of one of the clusters:

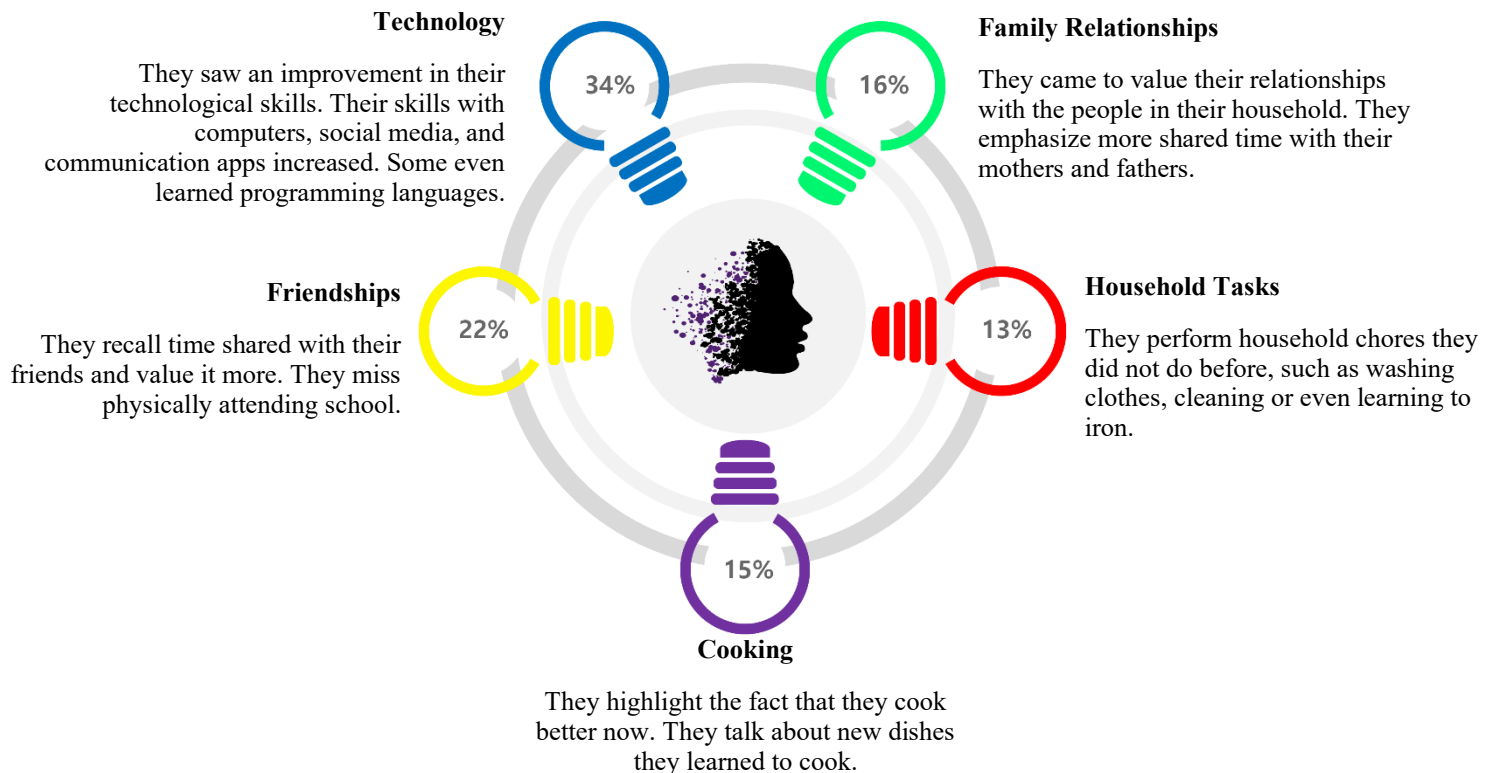


For each of the five clusters we have a heat matrix, with its top 25 words calculated using NMF analysis.

You can adjust the number of words in the 25 word list depending on whether the results obtained are conclusive enough or not. That is, if more information is needed to get a more accurate understanding of the results, you can increase the number of words in the list. On the other hand, if the results are too detailed and you want to simplify, you can decrease the number of words in the list.

4. RESULTS

In all of the preceding steps, we can observe the internal functioning of the algorithm and determine its outputs. The outcome is evident: there are five separate segments in our sample and each segment addresses distinct subjects in their responses. The primary subjects mentioned by each segment are distinctively recognized through the use of Non-Negative Matrix Factorization (NMF) and correlation matrices. The results are presented graphically as follows:



5. CONCLUSION

By combining descriptive statistics and machine learning methodologies, this algorithm offers the ability to convert words into numbers, segment respondents, and generate phrases that explain what the respondents are talking about in each segment.

Overall, this tool offers great promise for the future of quantitative research and provides valuable insights for decision-makers.

The results from the study conducted by KNACK for UNICEF in 2020 demonstrate the potential of the tool to provide clear and insightful results. Furthermore, using this tool would save researchers a significant amount of time and effort that would otherwise have to be spent on manual coding and analysis. With the efficiency and speed of the algorithm, researchers are able to get more accurate results in a matter of minutes, freeing up time for further analysis and interpretation of the results.

6. RECOMMENDATIONS FOR FUTURE RESEARCH

It is important to note that we are still in the early stages of this field and there is a lot of room for improvement. This technology is constantly evolving, and new developments are being made all the time. Therefore, while this tool may be advanced today, it is by no means the final product. With further advancements, it will likely become even more sophisticated and efficient in the future.

Another recommendation for future tests with this tool is to try it with the greatest possible number of cases. By increasing the number of cases, we will obtain a larger quantity of responses. We are familiar with the results obtained with a few thousand cases, but we are unaware of the tool's capabilities when dealing with hundreds of thousands of responses.

It is also worth testing the tool with less specific questions. Segmenting responses according to different opinions regarding a specific topic is straightforward with specific questions. However, what happens if we use it with samples that may be discussing completely different matters?



Federico Adroque

7. LIBRARIES REFERENCES

A. NLTK: <https://www.nltk.org/>

B. Scikit-Learn: <https://scikit-learn.org/>

1. TF-IDF: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
2. PCA: <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>
3. KMeans: <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>
4. NMF: <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html>

MANAGERIAL AND ACADEMIC CONSIDERATIONS FOR THREE APPROACHES TO WILLINGNESS TO PAY

BRYAN ORME

KEITH CHRZAN

SAWTOOTH SOFTWARE

GREG ALLENBY

THE OHIO STATE UNIVERSITY

BACKGROUND AND INTRODUCTION

Researchers in marketing sciences and economics have for over forty years used conjoint analysis¹ to estimate respondents' preferences (utilities) for brands, features, and prices. When the conjoint analysis study involves a price attribute, researchers have applied various approaches to estimate respondents' willingness to pay (WTP) for brands or feature enhancements.

We review three approaches to estimating WTP: 1) Algebraic approach applied to conjoint utilities (Swait et al. 1993, Orme 2001), 2) a Market Indifference approach that relies on market simulations (Orme 2001, Orme 2021), and 3) a Social Surplus approach that uses part-worth utilities and accounts for the competitive set (Allenby et al. 2014). The first (Algebraic) approach does not consider the strength of the firm's offering or its standing relative to competitors. The second and third approaches are affected by the strength of the firm's offering as well as the strength of the competitive set, which may include the outside good (i.e., the None alternative).

We explore characteristics of the three WTP approaches in the context of a realistic conjoint analysis study on vacation preferences and give recommendations for practice. We find little to recommend regarding the traditional Algebraic approach. We recommend the Market Indifference via market simulations and the Social Surplus approaches, with the choice depending on the focus of the researcher and business problem at hand. If the firm's perspective of restoring market share (unit volume) is paramount, then we prefer the Market Indifference simulation approach. We prefer the Social Surplus approach when the emphasis is on restoring utility to the consumer due to the loss of a feature or loss in a preferred firm's competitive stature due to infringement of a patent. Social Surplus WTP has been accepted in the courts in cases involving patent infringement.

EXAMPLE DATASET

We illustrate the three WTP methods using a CBC (Choice-Based Conjoint) example representative of those commonly applied in practice. We fielded our CBC study in 2016 among 600 respondents involving choice of different cruise vacation offers as described on six attributes:

¹ Also known as Discrete Choice Experiments, Tradeoff analysis

- **Destination:** Mexican Riviera, E. Caribbean, W. Caribbean, Alaska, Norway, Mediterranean
- **Cruise line brand:** Norwegian, Disney, Royal Caribbean, Princess, Holland America, Carnival
- **# Days:** 7, 8, 9, 10, 11
- **Stateroom:** Inside stateroom, Ocean View stateroom, Ocean View Balcony
- **Age of Ship:** Older, Newer
- **Price per person per day:** \$100, \$125, \$150, \$175, \$200

Each respondent completed 15 choice tasks, where each choice task consisted of 4 concepts. The design, fieldwork, and analysis all used Sawtooth Software's Lighthouse Studio platform. Although our CBC study did not include a None alternative, most CBC studies would include a None. However, the fact that this CBC dataset did not include a None option does not alter our findings or recommendations regarding the three WTP methods covered here.

We estimated respondent preferences using a hierarchical Bayes Multinomial Logit (HB-MNL) model, with the price effect constrained to be negative.

DEFINING THE THREE WTP METHODS

Algebraic Approach

The Algebraic approach is a common WTP method that has been applied since the 1990s (Swait et al. 1993, Orme 2001) and likely extending earlier to the 1980s. Given a conjoint analysis experiment that involves product features and price, we can calculate WTP for any non-price attribute level in the study compared to another level within the *same* attribute.

It is calculated given the formula:

$$W = (U' - U) / P$$

Where:

W = Willingness to Pay

U' = utility for the firm's enhanced product

U = utility for the firm's base case product

P = utiles per dollar (typically from a linear price function)

We typically use utilities estimated at the individual level in the formula, such as from HB-MNL (often with the linear price coefficient constrained negative). Thus, we calculate WTP for each respondent. Some respondents may have very small estimates of utiles per dollar and the WTP values can become extremely large, extrapolating well beyond the price range included in the experiment. To stabilize estimates of WTP in the face of extreme outliers, we typically take the median of W across respondents.

Market Indifference Simulation Approach

The Market Indifference (MI_WTP) simulation approach relies on first creating a market (choice) simulator. The market simulator may be built in a variety of ways, such as from individual-level utilities (HB draws or point estimates), segment-based utilities (e.g., latent class MNL), or pooled MNL estimation. The underlying utility model may involve main effects, alternative-specific effects, interaction terms or cross-effects. The choice rule can be first choice, share of preference, randomized first choice, or any number of choice rules wherein respondents “vote” on competitive offerings in the market simulation scenario and the shares of votes (shares of preference) sum to 100%.

The MI_WTP approach involves simulating market choice (share of preference) for the firm’s offering (both base case and enhanced) when it is placed in competition with other competitors’ offerings and (often) the None alternative. The MI_WTP steps are as follows:

1. Run a market simulation involving the firm’s original (not enhanced) alternative at price P_0 along against competitors and (typically) the None. As usual, the simulated shares of preference sum to 100% across alternatives. Record the firm’s base case share of preference (S_0) given P_0 .
2. Run a new market simulation where the firm’s alternative is enhanced, holding competitors constant. Record the firm’s new share of preference (S_1) at P_0 , which we expect to be higher than S_0 due to the product enhancement.
3. Using a manual or automated search process, find a new global price P_1 applied to the firm’s enhanced alternative such that S_1 is returned to the original S_0 .

Note that the MI_WTP approach is not computed at the individual level, even though the shares of preference typically result from the accumulated “votes” of individual respondents stemming from their individual-level part-worth utilities.

Social Surplus Approach

The Social Surplus approach (SS_WTP, Allenby et al. 2014) involves an algebraic closed-form expression that has some similarities to the Algebraic approach as previously defined. However, it is more complete, explicitly accounting for competitors and the None alternative.

Given competitors b and c and None alternative n, we typically compute WTP for the firm’s product a at the individual level as follows:

$$W = [\text{LN} (A' + B + C + N) - \text{LN} (A + B + C + N)] / P$$

Where:

W = Willingness to Pay

A' = exponentiated² utility for the firm’s enhanced product

B = exponentiated utility of competitor b

C = exponentiated utility of competitor c

N = exponentiated utility of the None alternative n

P = utiles per dollar (typically from a linear price function)

² The syntax in Excel for exponentiating the utility of a product alternative is =EXP(U_i) where U_i is the total utility for the i^{th} product alternative.

We typically apply the SS_WTP formula to part-worth utilities estimated at the individual level, such as from HB-MNL (often with linear price coefficient constrained negative). Thus, we calculate SS_WTP for each respondent for a given feature enhancement. Some respondents may have very small estimates of utiles per dollar and the SS_WTP values can become extremely large, extrapolating well beyond the price range included in the experiment. To stabilize estimates of SS_WTP in the face of extreme outliers, we typically take the median of W across respondents.

TYPICAL MAGNITUDE OF WTP MEASURES

Previous research has found that the MI_WTP is often 10%–20% lower than the traditional Algebraic approach (Orme 2021, Moore and Bhudiya 2022). However, this depends on the strength of the firm's offering relative to the competition.

In practice, the SS_WTP approach is usually lower than both the algebraic and MI_WTP approaches. In the corner case where the firm's initial offering has a very high share of preference (e.g., >99%) the SS_WTP essentially matches the WTP from the traditional algebraic approach. In the corner case where the firm's initial offering has a very low share of preference (e.g., <1%) SS_WTP approaches zero.

WTP GIVEN STRENGTH OF FIRM'S OFFERING

The traditional Algebraic WTP approach is invariant to the strength of the firm's offering relative to competitors. It effectively assumes a monopoly wherein each respondent can only obtain the product enhancement from the firm and is forced to pay their estimated WTP to obtain it (and cannot walk away by selecting the None alternative).

The MI_WTP and SS_WTP are both affected by the firm's position relative to the competition as well as the attractiveness of the None alternative. To illustrate this, we configure a market scenario involving the cruise line choice dataset we described previously. The firm's base case product initially is set as the Carnival brand at \$100 per night. Three competitors are also specified in the competitive scenario. In this base case scenario, the firm's initial share of preference is 48%. Given this competitive scenario, the WTP for the Norwegian brand over Carnival is as follows for the three WTP approaches:

Exhibit 1:
WTP for Norwegian over Carnival Brand
When Firm's Product Share of Preference is 48%

Algebraic	\$17.58
MI_WTP	\$17.60
SS_WTP	\$3.15

(While MI_WTP is usually 10–20% lower than the Algebraic approach,
this is not the case for this specific example.)

Next, we decrease the quality of the firm's offering by giving it worse levels on certain attributes such as lower quality stateroom and an older ship, while increasing the quality of the competitors' offering by improving their features. The new share of preference for the firm's product is 12% and the new WTP values for Norwegian brand over Carnival are:

Exhibit 2:
WTP Norwegian over Carnival Brand
When Firm's Product Share of Preference is 12%

Algebraic	\$17.58
MI_WTP	\$4.70
SS_WTP	\$0.23

As we mentioned before, the Algebraic approach is blind regarding the strength of the firm's offering relative to the competition. The other two WTP approaches consider and are sensitive to the different competitive context. The MI_WTP approach sees its WTP for Norwegian over Carnival decrease from \$17.60 (Exhibit 1) to \$4.70 (Exhibit 2) when the firm's product is made weaker and pitted against stronger competition. The SS_WTP approach is even more sensitive, with the Norwegian brand decreasing from the original \$3.15 premium (Exhibit 1) to \$0.23 (Exhibit 2) when the firm's product is weaker and the competition is stronger. Indeed, it is more realistic for the firm to experience lower WTP for an enhancement when the market is less enthusiastic about its offering to begin with in both an absolute sense and relative to competitors.

WTP GIVEN NUMBER OF ASSUMED COMPETITORS

The Algebraic approach does not consider competition, so it is invariant over the number of assumed competitors in the marketplace.

The MI_WTP and SS_WTP are both sensitive to the number of competitors, as both approaches are affected by competition. To assess the effect of competitors on WTP for these latter two approaches, we employed a Sampling of Scenarios (SOS) approach (Orme 2021). In the SOS approach, we run the market simulation hundreds of times, where for each simulation we draw random characteristics (and prices) for the competitive products. We compute the WTP as previously described for MI_WTP and SS_WTP for each of the hundreds of simulation scenarios and take the median WTP result. This generalizes our WTP findings beyond the particular configurations of competitors that we might select.

As in the previous section, we compute the WTP for the Norwegian brand over Carnival. We now vary the number of randomly-drawn competitors in the market simulations from 1 to 49. The WTPs under the different number of competitors for MI_WTP and SS_WTP are as follows:

Exhibit 3:
WTP Norwegian over Carnival Brand
by Different Numbers of Competitors

#Competitors	1	3	9	14	19	49
MI_WTP	\$20.80	\$16.96	\$14.16	\$12.82	\$12.50	\$12.60
SS_WTP	\$5.53	\$1.91	\$0.33	\$0.17	\$0.13	\$0.06

For this data set, once there are 14 or more competitors, MI_WTP is quite stable. We tested this out to 200 assumed competitors and continued to see MI_WTP of around \$12 to \$13. For the MI_WTP approach, somewhere from 10 to 20 assumed competitors would seem to strike a balance between computational speed and stability of the results. In contrast, SS_WTP continues to decrease as more competitors are assumed. Given enough competitors, SS_WTP approaches zero. For the SS_WTP approach, we generally recommend setting the competitive scenario to mimic either the number of products in the marketplace or the number of alternatives used in the choice questionnaire.

WTP GIVEN MULTIPLE ENHANCEMENTS

Practitioners often find that their clients (whether internal or external) are tempted to add WTP values across features, extrapolating the total WTP across multiple feature enhancements. WTP calculations are estimated assuming one product enhancement at a time. To treat them simply as additive ignores the likely possibility of satiation.

For the examples in this section, we modified the individual-level utilities to add new attribute levels with corresponding fixed utility value enhancements to the original dataset. We did this so we could investigate three new feature enhancements that would have exactly +0.5 utiles of improvement for every respondent in the dataset over the base case product. This allowed us to examine how the three WTP approaches dealt with the multiple constant utility feature improvements. Given what we know about economics and consumer psychology, we'd naturally expect there to be diminishing marginal effects for multiple equal enhancements due to satiation.

For the algebraic approach, each additional 0.5 utile product improvement leads to a constant additional monetary increase in WTP. The algebraic approach assumes linearity of WTP for the three equal 0.5 utility improvements: 3x the utility improvement leads to 3x the WTP.

For the MI_WTP approach, three equal utility improvements led to WTP nearly linear in response to the three improvements in utility for the firm's offering. In practice, most Sawtooth Software users do not fit a linear price function when building market simulators. Thus, we decided to create a new version of the simulator that fit a standard part-worth (effects-coded) function for price. Using this second simulator built including a non-linear price function, the MI_WTP shows diminishing marginal WTP for three equal utility step improvements (Exhibit 4).

For the SS_WTP approach, equal increment improvements to the firm's product always lead to accelerating marginal WTP returns (i.e., more than the sum of individual components) (see Exhibit 4). More details regarding this result are provided in Appendix A.

Exhibit 4:
WTP Given Three Equal 0.5 Utility Improvements

	1 Improvement	2 Improvements	3 Improvements
MI_WTP (linear Price)	\$22.98 (1x)	\$45.42 (1.98x)	\$68.22 (2.97x)
MI_WTP (nonlinear Price)	\$22.49 (1x)	\$36.30 (1.6x)	\$49.38 (2.2x)
SS_WTP	\$15.03 (1x)	\$35.74 (2.4x)	\$62.62 (4.2x)

One of the authors questioned whether MI_WTP with the linear specification of price will always show linear or potentially slightly diminishing WTP returns given multiple equal utility improvements. We tested different market simulation conditions (strength of firm's product vs. competition) and did find a case which reflects accelerating WTP returns. So, it isn't always guaranteed that the MI_WTP approach will lead to linear or diminishing marginal effects of utility on WTP.

PRACTICAL CONSIDERATIONS

Our experience as consultants is that clients usually like to have market simulators delivered as an Excel file. When they request WTP, they typically like to see it reported on a separate tab in their Excel simulator or reported as additional stubs (row variables) within cross-tabs. The algebraic and SS_WTP approaches are very easy to implement in Excel and to report by user-

selected segment. The MI_WTP is more complex to implement, as it involves iteratively solving for the price increase that drives the share of preference for the firm's product back to its base case share. To report MI_WTP by segment, a separate search is needed for the market indifference price for each segment. Sawtooth Software has implemented the SS_WTP approach within its desktop market simulator, but those writing their own market simulation tools might find it harder to implement MI_WTP than the algebraic or SS_WTP approaches.

STRENGTHS AND WEAKNESSES SUMMARY

	Algebraic	MI_WTP	SS_WTP
Ease of Computation:	Very easy, can be done in Excel	Harder, requires market simulator and iterative search	Very easy, can be done in Excel
Affected by Competitors:	No, ignores them	Yes	Yes
Magnitude of WTP:	Larger than MI_WTP and SS_WTP	Typically about 10%–20% lower than Algebraic Approach	Typically much lower than Algebraic and MI_WTP
Sensitivity of WTP to firm's utility and brand equity:	None, doesn't consider it	Sensitive	Extremely sensitive
Sensitivity of WTP to number of assumed competitors:	None, doesn't consider it	Moderately sensitive until about 10–20 competitors, somewhat stable thereafter	Extremely sensitive, can drive WTP to near \$0 with enough assumed competitors
How WTP affected by multiple simultaneous feature improvements:	Assumes strict additivity of WTP across features	Depends on the data set and the shape of the price function	Reflects accelerating marginal WTP for additional feature improvements

RECOMMENDATIONS FOR CONSULTING PRACTICE

We do not recommend the standard Algebraic approach. It pays no attention to the quality of the firm's offering nor to the strength of competition. The estimated WTP values are often too high and unrealistic. The difference between the MI_WTP and SS_WTP measures is due to what is being restored in the calculations. For MI_WTP, shares are being restored, while for SS_WTP

consumer utility is being restored. The MI_WTP measure focuses on one of the brands in this restoration (e.g., the Carnival brand), and the shares of competing brands are not factored into the MI_WTP valuation. That is, the MI_WTP calculation does not seek to restore all of the shares to their original level, just the focal brand.

The SS_WTP measure seeks to restore consumer utility, which is affected by all of the choice alternatives. The multinomial logit model of choice used in choice-based conjoint analysis is a variant of a random utility model, where analysis predicts the probability of choice. Respondents are assumed to be utility maximizers, and respondent welfare is measured in terms of the expected maximum utility from a choice scenario. Since each choice alternative is measured with error, and each has some probability of being chosen, the calculation of SS_WTP described above involves all choice alternatives. If a feature enhancement is performed on a weak brand, it will lead to a small estimate of SS_WTP because the improvement will likely leave the weak brand relatively weak and not have a large effect on expected maximum utility. If the enhancement is to a strong brand, the SS_WTP will be larger because stronger brands have a higher probability of being selected and make a greater contribution to expected maximum utility. If the enhancement is made in a market with many competitors, then the enhancement to any one brand will be dampened.

The question to consider when choosing between MI_WTP and SS_WTP is what should be restored – share or utility? Firms will often want to focus on the sales of their own offering and care most about the revenue it generates. In this case, MI_WTP would be a better measure of assessing the value of a product enhancement. SS_WTP would be preferred when addressing questions dealing with whether consumers are better off because of the enhancement. These questions are important when measuring societal gains and harm, such as determining the value of product mislabeling that affects consumers because their purchases are inappropriately drawn to the brand making false claims, or cases of patent infringement where harm plays itself out across all the brands in a market. The value of enhancing public good, such as a park, a stream or a lake, is also better measured in terms of SS_WTP because the goal is to enhance the lives of people, not to increase usage of any one specific resource. So, while SS_WTP takes a more encompassing view of value, it may not be preferred when analysis is focused on a specific brand or offering. Both MI_WTP and SS_WTP have their place in analysis.



Bryan Orme



Keith Chrzan



Greg Allenby

APPENDIX A

Accelerating Marginal WTP for SS_WTP

For the examples below, we assume the firm's initial utility is 0. We also assume each competitor in the market scenario has a utility of 0.

Assuming One 0.5 Utile Improvement to the Firm:

$$SS_WTP = [\text{LN} (\exp(0.5) + \exp(0) + \exp(0) + \exp(0)) - \text{LN} (\exp(0.0) + \exp(0) + \exp(0) + \exp(0))] / 0.01$$

$$SS_WTP = [\text{LN} (1.65 + 1 + 1 + 1) - \text{LN} (1.00 + 1 + 1 + 1)] / 0.01$$

$$SS_WTP = \$15.03$$

(A Single 0.5 utile improvement leads to \$15.03 increase in WTP)

Assuming Two 0.5 Utile Improvements to the Firm:

$$SS_WTP = [\text{LN} (\exp(1.0) + \exp(0) + \exp(0) + \exp(0)) - \text{LN} (\exp(0.0) + \exp(0) + \exp(0) + \exp(0))] / 0.01$$

$$SS_WTP = [\text{LN} (2.72 + 1 + 1 + 1) - \text{LN} (1.00 + 1 + 1 + 1)] / 0.01$$

$$SS_WTP = \$35.74$$

(Two 0.5 utile improvements leads to \$35.74 increase in WTP; whereas if additive we expect \$15.03 x 2 = \$30.06)

Assuming Three 0.5 Utile Improvements to the Firm:

$$SS_WTP = [\text{LN} (\exp(1.5) + \exp(0) + \exp(0) + \exp(0)) - \text{LN} (\exp(0.0) + \exp(0) + \exp(0) + \exp(0))] / 0.01$$

$$SS_WTP = [\text{LN} (4.48 + 1 + 1 + 1) - \text{LN}(1.00 + 1 + 1 + 1)] / 0.01$$

$$SS_WTP = \$62.62$$

(Three 0.5 utile improvements leads to \$62.62 increase in WTP; whereas if additive we expect \$15.03 x 3 = \$45.09)

SS_WTP for two or more simultaneous improvements is always greater than the sum of their individual WTPs (as long as all are positive).

REFERENCES

- Allenby, Greg, Jeff Brazell, John Howell, and Peter Rossi (2014), “Economic Valuation of Product Features.” *Quantitative Marketing and Economics* 12:421–456.
- Moore, Chris and Manjula Bhudiya (2022), “An Empirical Comparison of Willingness to Pay Methods.” *Proceedings of the Sawtooth Software Conference*, pp 85–100, Provo, UT.
- Orme, Bryan (2001), “Assessing the Monetary Value of Attribute Levels with Conjoint Analysis: Warnings and Suggestions.” *Sawtooth Solutions*, Sequim WA.
- Orme, Bryan (2021), “Estimating Willingness to Pay (WTP) Given Competition in Conjoint Analysis.” *Proceedings of the Sawtooth Software Conference*, pp 125–140, Provo, UT.
- Swait, Joffre, Tulin Erdem, Jordan Louviere, and Chris Dubelaar (1993), “The Equalization Price: A measure of consumer-perceived brand equity.” *International Journal of Research in Marketing*, 10 (1993) 23–45.

SWIPE RIGHT ON SIMPLICITY: EXAMINING THE THEORETICAL AND PRACTICAL VIABILITY OF CHOICE SETS OF SIZE ONE

JEFFREY P. DOTSON

JOHN HOWELL

MARC DOTSON

BRIGHAM YOUNG UNIVERSITY

CRAIG LUTZ

QUALTRICS

INTRODUCTION

In recent years, the mobile user experience has become an essential aspect of conducting surveys and studies. As businesses increasingly rely on data-driven decision-making, it is crucial to ensure that survey participants are engaged and motivated to provide accurate, high-quality responses. This paper proposes an approach to addressing the challenges associated with traditional conjoint analysis on mobile devices, enhancing the user experience and potentially improving data quality.

Conjoint analysis is a widely used research method for understanding consumer preferences and informing business decisions. However, the traditional multi-alternative grid-based format often proves cumbersome and unengaging for respondents using mobile devices, leading to a less-than-optimal user experience. This research explores the impact of mobile user experience on data quality and investigates the potential benefits of a single-alternative interface for conjoint studies conducted on mobile devices. This intuitive, engaging, and gamified interface seeks to improve the user experience and increase the overall quality of the collected data.

The proposed solution involves the implementation of a user-centric data collection technique inspired by the Tinder-style interface found in online dating apps. In this approach, respondents are presented with a single product profile characterized by brand, price, and various features. Participants are then asked if they would consider purchasing the product, swiping right for a positive response and left for a negative one. This process is repeated multiple times, allowing for the collection of preference data in a more engaging and user-friendly manner. In assessing the viability of this approach, we consider the following 4 research questions:

1. What is the impact of multi-alternative choice sets on data collected via mobile devices?
2. What is the impact of using single-alternative choice sets on data quality and the subjective user experience?
3. What is the impact of using single-alternative choice sets on statistical information? Can fewer alternatives be shown while still gathering the same amount of statistical information as traditional multi-alternative conjoint studies?
4. How do current mobile data collection approaches for conjoint studies perform, particularly carousel grids and stacked cards?

To answer these research questions, we conducted three studies: two empirical studies and one simulation study.

Study 1: An empirical study comparing single-alternative and multi-alternative grid choice tasks on both mobile devices and desktop computers. This study aims to address research questions 1 and 2.

Study 2: A simulation study to determine the impact of single-alternative choice tasks on statistical information, addressing research question 3.

Study 3: Another empirical study, focusing on mobile devices and comparing different mobile-centric approaches, including grid, carousel, and stacked card approaches, as well as the single-alternative approach. This study aims to address research question 4.

Taken collectively, these studies suggest that the proposed single-alternative conjoint approach may be a reasonable way to improve the quality of conjoint data collected on mobile devices. We find that it is more enjoyable, quicker to complete, and yields results that are consistent with data collected on large-screen devices. Further, we find that it performs at least as well (if not better) than current approaches that have been optimized for mobile devices (i.e., locking carousel and stacked cards). Although additional research is warranted, we feel optimistic about the potential value of this approach.

Details of our study appear below:

STUDY 1: ASSESSING DATA QUALITY AND USER EXPERIENCE OF SINGLE-ALTERNATIVE CHOICE SETS COMPARED TO TRADITIONAL MULTI-ALTERNATIVE GRIDS

Study 1 aims to address two key research questions: How severe is the data quality issue in multi-alternative choice sets on mobile devices, and does the proposed single-alternative choice set solution effectively improve data quality and user experience? We use a conjoint study for barbecue sauce preference to investigate these questions, comparing single-alternative and multi-alternative grid choice tasks on both mobile devices and desktop computers.

Methodology

This study implements a 2 X 2 experimental design that crosses design type (single- vs. multi-alternative) against device type (desktop vs. mobile). We worked with a panel provider to fill each of the 4 cells. Respondents were broadly invited to participate in the study. Respondent device type was measured (both directly and indirectly) and subjects were assigned to a cell accordingly. This may create the conditions for selection (e.g., younger respondents are more likely to use mobile devices and may be disproportionately represented in those conditions). Demographic data was measured, but still needs to be analyzed to refute this hypothesis. In total, data were collected from 446 respondents from a commercial panel, divided across four conditions:

- Desktop conjoint with a multi-alternative grid (three alternatives plus a no-choice option).
- Desktop conjoint with a single-alternative choice set.
- Mobile conjoint with a multi-alternative grid (three alternatives plus a no-choice option).
- Mobile conjoint with a single-alternative choice set.

The primary focus is the comparison between the multi-alternative grid and single-alternative approaches on mobile devices. Screenshots of each condition appear in Figure 1:

Figure 1:
Screenshots of the 4 Conditions Used in Study 1

Desktop: 3 + No Choice

Desktop: 1 + No Choice

Mobile: 3 + No Choice

Mobile: 1 + No Choice

Evaluation Metrics

The following evaluation metrics will be used to assess the quality of the proposed approach:

- Response time: Average time taken by respondents to complete the study under different approaches.
- Subjective evaluation: Measures of the user's experience with the survey, such as ease of completion and enjoyment.
- Parameter similarity: Assessing whether the different approaches yield similar results.

Results

Study results appear in Table 1. The key findings of Study 1 are as follows:

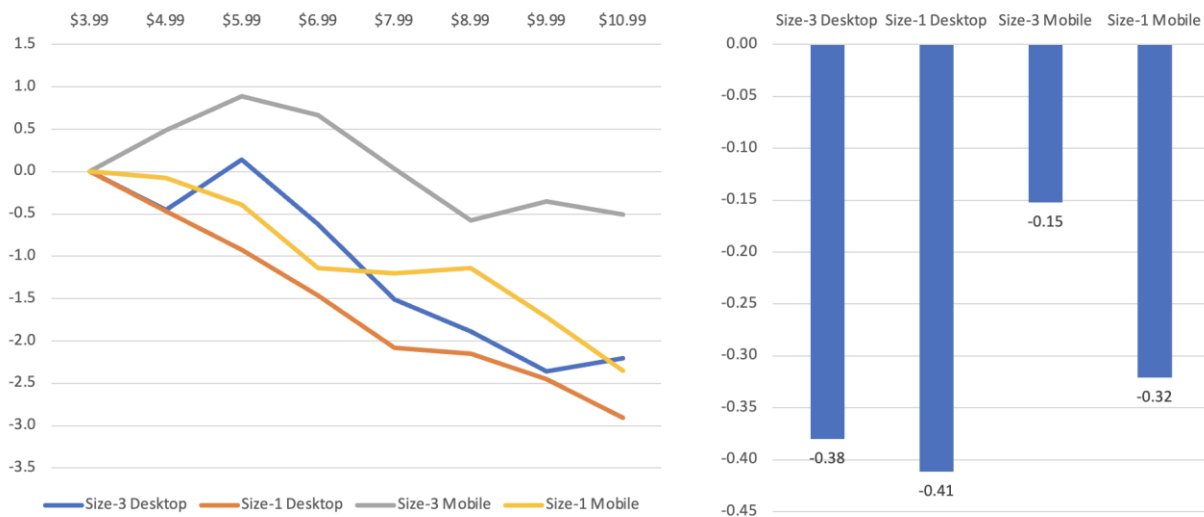
Table 1:
Evaluation Metrics for Study 1

	Desktop Grid	Desktop Single	Mobile Grid	Mobile Single
Response Time	2.5	3.3	2.4	2.8
Easy	64%	68%	62%	69%
Enjoyable	40%	43%	40%	45%
Recovery (MSE)	Base	0.23	0.85	0.23

- *Response time*: Single-alternative choice tasks took slightly longer to complete compared to the grid-based tasks on both desktop and mobile devices. However, the response time difference was not substantial.

- *Subjective evaluations:* Respondents found the single-alternative choice tasks easier to complete and more enjoyable than the grid-based tasks
- *Parameter similarity:* The single-alternative choice tasks on both desktop and mobile devices yielded similar results to the gold standard. However, the grid-based tasks on mobile devices resulted in poorer parameter similarity, particularly for price coefficients. This is illustrated in Figure 2. The price curves and estimated linear price coefficient are similar for all approaches with the exception of the Size-3–Mobile condition. This condition dramatically understates price sensitivity relative to the other approaches.

Figure 2:
Estimated Price Parameters—Partworths (left) and Linear (right)



Conclusion

Study 1 concludes that the grid-based format on mobile devices produces poor data quality as expected. The single-alternative choice sets show promising results, providing comparable data quality and improved user experience compared to the traditional grid-based format. These findings support the potential benefits of implementing single-alternative choice sets in conjoint studies conducted on mobile devices.

STUDY 2: ASSESSING THE NUMBER OF CHOICE TASKS REQUIRED FOR EQUIVALENT INFORMATION BETWEEN SINGLE- AND MULTI-ALTERNATIVE CONJOINT DESIGNS

Study 2 aims to determine the number of choice tasks required to obtain equivalent information between single-alternative and multi-alternative data collection techniques. The study conducts a simulation experiment considering three conjoint designs: one with standard attributes and levels, one with many attributes, and one with many levels. Details of the simulation study design appear in Table 2.

Table 2:
Simulation Study Design

	Name	# Attributes	# Levels	Continuous	Total
--	------	--------------	----------	------------	-------

				Price	Parameters
Scenario 1	Standard Design	6	$4 \times 3 \times 2^3 \times 1$	Yes	10
Scenario 2	Many Attributes	12	$3^4 \times 2^8$	No	17
Scenario 3	Many Levels	8	$15 \times 6 \times 2^5 \times 1$	Yes	26

Methodology

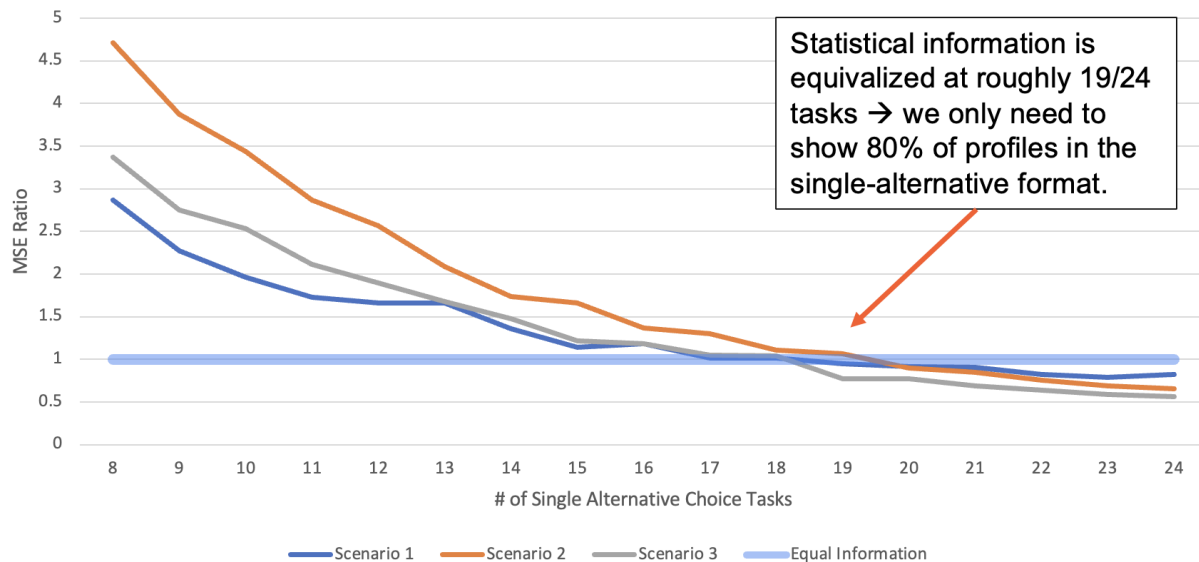
The simulation follows these steps:

1. Simulate data with known true part-worths.
2. Estimate a Hierarchical Bayes Multinomial Logit Model (HBMNL) using a subset of single-alternative choice tasks.
3. Compute the Mean Squared Error (MSE) for the estimated part-worths relative to the true values.
4. Repeat the process, adding one more single-alternative choice task each time.
5. Replicate this process 40 times for each design and summarize the results.

The point of equivalent information is determined by comparing the MSE ratio between single-alternative and multi-alternative choice sets. Results for each scenario and number of choice tasks are plotted in Figure 3.

The results show that approximately 19 out of 24 single-alternative choice tasks are required to achieve equivalent information between single- and multi-alternative choice sets. This indicates that only about 80% of profiles need to be shown in the single-alternative format to obtain the same statistical information as an 8-task design with 3 alternatives per task.

Figure 3:
MSE Ration (single-alternative vs. multi-alternative with all tasks) for various scenarios



Conclusion

Study 2 reveals that the information function exhibits concavity, implying that it is not necessary to match the number of concepts in single-alternative and multi-alternative designs for equivalent statistical information. Presenting approximately 80% of single-alternative choice tasks is sufficient to achieve equivalent statistical information. This finding suggests that fewer questions can be asked without sacrificing statistical efficiency, potentially reducing respondent time and effort.

STUDY 3: ASSESSING THE PERFORMANCE OF MOBILE DATA COLLECTION APPROACHES FOR CONJOINT ANALYSIS

Study 3 evaluates the performance of current mobile data collection approaches for conjoint analysis relative to the proposed approach. The goal is to understand the efficiency, ease of use, and subjective experience of various mobile interfaces in conjoint studies. The study implements a swipeable interface and compares it with traditional grid, carousel, and stacked cards approaches.

Methodology

A conjoint study was designed for mattresses using brand, price, and various product attributes. Data was collected from 1,000 respondents, who completed the study on mobile devices across four conditions:

1. Grid—All K alternatives are shown on the screen. Users must scroll horizontally (and vertically) to examine them.
2. Locking carousel grid—1 of K alternatives are displayed on the screen at a time. Users rotate through alternatives by swiping horizontally. This is the mobile approach used by Sawtooth.
3. Stacked cards—1 of K alternatives are displayed on the screen at a time. Users rotate through alternatives by swiping vertically. This is the mobile approach used by Qualtrics.
4. Singleton (swipeable)—Users are shown a single alternative in each choice task. They swipe right if they would buy it and left if they would not.

For multi-alternative designs (options 1–3 above), respondents were shown four alternatives and a “no choice” option across 12 choice tasks. In the single-alternative design, respondents were shown 48 swipeable profiles. Like Study 1, Study 3 measured choice, response time, and subjective experience with the survey. Results appear in Table 2.

Table 3:
Evaluation Metrics for Study 2

	Grid	Carousel	Stacked	Swipe
Response Time	5.3	5.6	5.4	4.4
Easy	74%	71%	74%	73%
Enjoyable	44%	47%	41%	62%
Recovery (MSE)	0.15	base	0.08	0.09

Summary of Results

- *Response time:* The swipeable condition was nearly one minute faster than the other approaches, despite having 48 choice tasks compared to 12 in the other conditions.
- *Ease of completion:* All conditions were rated similarly in terms of ease of completion.
- *Enjoyment:* The swipeable condition was rated significantly higher in enjoyment than the other conditions.
- *Parameter recovery:* The swipeable, carousel, and stacked formats showed similar Mean Squared Error (MSE) values, while the grid approach had about twice the MSE, indicating greater differences and errors.

Conclusion

Study 3 reveals that the swipeable conjoint task was quicker to complete and more enjoyable for respondents compared to the grid, carousel, and stacked card approaches. Additionally, the results between the swipeable, carousel, and stacked formats were more similar to each other than to the grid format. These findings suggest that implementing more intuitive and user-friendly mobile interfaces can lead to better user experience and data quality in conjoint studies.

CONCLUSIONS AND NEXT STEPS

This research emphasizes the significance of the user experience and interface design in conjoint studies. There is strong evidence that multi-alternative grid questions are not effective on mobile devices, yielding poor data quality and parameter recovery. Singleton choice sets appear to be a better option for mobile data collection, as they are more enjoyable and quicker to complete than current approaches like locking carousels, stacked conjoint, and multi-alternative grids.

To strengthen this view, there are a variety of open questions that need to be addressed including:

- Developing better measures of respondent quality.
- Improving parameter recovery: Examining how to define “truth” relative to observed data and explore the impact of different data collection modalities on respondent psychology and data quality.
- Addressing the outside good: Consider ways to reinforce the definition of the outside good thus facilitating more consistent comparisons.
- Identifying boundary conditions for the swipe format: Are there conditions when we would expect this approach to perform better/worse? For example, we believe this approach will work exceptionally well if the goal is to learn about product consideration. However, if the goal is to understand substitution patterns and market structure, we believe it is necessary to use multi-alternative choice.

Code, and data related to this project are available at:

<https://github.com/statuser/ChoiceSetsSizeOne>



Jeffrey P. Dotson



John Howell



Marc Dotson



Craig Lutz

FINDING CONTRASTIVE MARKET SEGMENTS WITH ARCHETYPAL ANALYSIS

JACOB NELSON
HARRIS POLL

I. INTRODUCTION

Market segmentation is valuable for understanding a market audience and gives marketers leverage over their market strategy. Traditionally, researchers use cluster analysis algorithms like K-Means and Latent class to identify market segments and these algorithms have been used with high success (Chrzan and White, 2021). However, the goals of the researcher are often inconsistent with the objectives of these algorithms. Cluster analysis seeks relatively homogenous segments, while researchers often don't care much about intra-segment homogeneity and are more concerned with finding contrastive differences between segments to exploit. While cluster analysis will often find differentiated segments regardless, it can miss important opportunities in the pursuit of segment homogeneity.

Archetypal analysis is an alternative segmentation approach more in line with research objectives where segments must be differentiated, but don't necessarily need to be homogenous and concise. This presentation will suggest adding this powerful approach to the practitioner's repertoire. Archetypal analysis was introduced by Cutler and Breiman in 1994 as a way of structuring data observations as convex combinations of extremal data values (Cutler and Breiman, 1994) and has been successfully adopted by many as a tool for market segmentation as well (Li et al., 2003).

This paper will demonstrate the advantages of adding archetypal analysis into the practitioner's toolkit. We will first briefly review the core objectives of market segmentation in practice. Then we will compare traditional cluster analysis techniques with archetypal analysis and show how they work to achieve the objectives of market segmentation differently. To help the researcher begin to adopt archetypal analysis, we provide a high-level explanation of how the algorithm works, as well as a brief tutorial on how to implement archetypal analysis in R. We'll then cover a few common questions about archetypal analysis and give advice on when to use this technique.

II. OBJECTIVES OF MARKET SEGMENTATION

Keith Chrzan, Senior Vice President at Sawtooth Analytics writes: "*Segmentation* helps marketers understand how *groups* of customers *differ* with respect to the products, messaging, or position that appeal to them. Understanding these differences gives marketers more *leverage* in designing or selling products to their customers" (Chrzan, 2021, emphasis added). Chrzan highlights for us that market segmentation is primarily about finding groups within a market audience that give marketers leverage in their market strategy. Segmentation can serve various facets of a market strategy, including identifying diverse customer needs, improving advertising effectiveness, finding niches in the market to reduce competition, optimizing resource allocation, and discovering profitable new opportunities.

Market research practitioners may use or see various criteria for evaluating the quality and usefulness of their market segmentations. These can be useful, but they also can distract the

practitioner from the broader goal of segmentation, that of finding groups within a market audience, the knowledge of which gives marketers leverage they are looking for in their market strategy.

Consider for instance, the practice of over-relying on goodness-of-fit statistics in various segmentation algorithms. Suppose there are two proposed segmentation solutions. The first solution fits the data exceptionally well, demonstrating a high degree of statistical fit. However, this solution also does not provide the market researcher with easily actionable insights or leverage to enhance their market strategy. In contrast, the second solution organizes the market audience in a narratively compelling way and offers clear and intuitive insights for market strategy. Although this solution may not fit the data as closely according to the segmentation algorithm (meaning that there might be some individuals within the market audience who do not fit perfectly into the defined segments), it provides valuable leverage for decision-making.

In such a scenario, which market segmentation solution would be preferable to the researcher? The latter solution, which offers meaningful insights and practical leverage for market strategy, is undoubtedly more valuable. In fact, the former solution, despite its superior fit to the data, may be considered useless if it fails to provide the necessary leverage for effective decision-making. This example illustrates that goodness of fit measurements are only useful to the extent that they enhance leverage, by improving segment differentiation and stability, thus enabling marketers to target these segments effectively. Therefore, it becomes evident that the primary objective of market segmentation is to maximize leverage over market strategy, with other principles of good market segmentation serving as supporting objectives.

III. TRADITIONAL MARKET SEGMENTATION WITH CLUSTER ANALYSIS: A FOUNDATION FOR COMPARISON

To better understand the advantages of archetypal analysis, we will first discuss how market segmentation is done traditionally with cluster analysis. This will help facilitate a meaningful comparison between cluster analysis and archetypal analysis later.

To perform a market segmentation, researchers will typically field a survey and collect data on their market audience, and then use a statistical or machine learning algorithm on select basis variables from that data to help them find the “best” or “most natural” groups according to some objective criteria. Most employed algorithms for segmentation belong to a class known as “cluster analysis.” In cluster analysis, the algorithm typically looks to identify groups that have the following two characteristics:

1. Similarity: Observations (respondents) found in collected data are similar to each other within group, and
2. Differentiation: Observations (respondents) found in collected data are different from each other across groups.

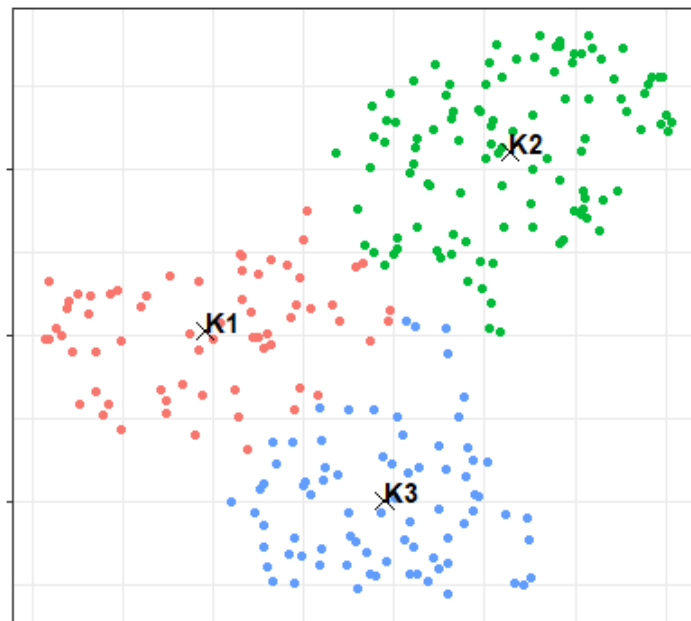
In essence, cluster analysis seeks to find segments that are homogenous, concise, and differentiated from each other. In this context, groups found in cluster analysis are called “clusters.”

One widely used algorithm for cluster analysis in segmentation is the “K-Means” analysis. Understanding K-Means analysis will provide a helpful contrast for later comparisons with archetypal analysis, so it is worth briefly reviewing this algorithm specifically. K-Means analysis uses artificial data points, called centroids, to identify and represent data clusters within a

dataset. It represents the center or average of a cluster of data. The objective of K-means analysis is to find the optimal positions to place these centroids in relation to the basis variables, minimizing the within-cluster sum of squares distances (i.e., the proximity of each observation to its centroid). Cluster membership is assigned based on which centroid is closest to each respondent or observation.

In the example depicted in Figure 1, two basis variables are plotted on the X and Y axes, and K-means analysis has positioned three centroids in the data per the researcher's specifications. Points K1, K2, and K3 represent the final converged positions of the centroids. Each observation in the data (represented by dots), is assigned to a cluster based on its nearest centroid.

Figure 1



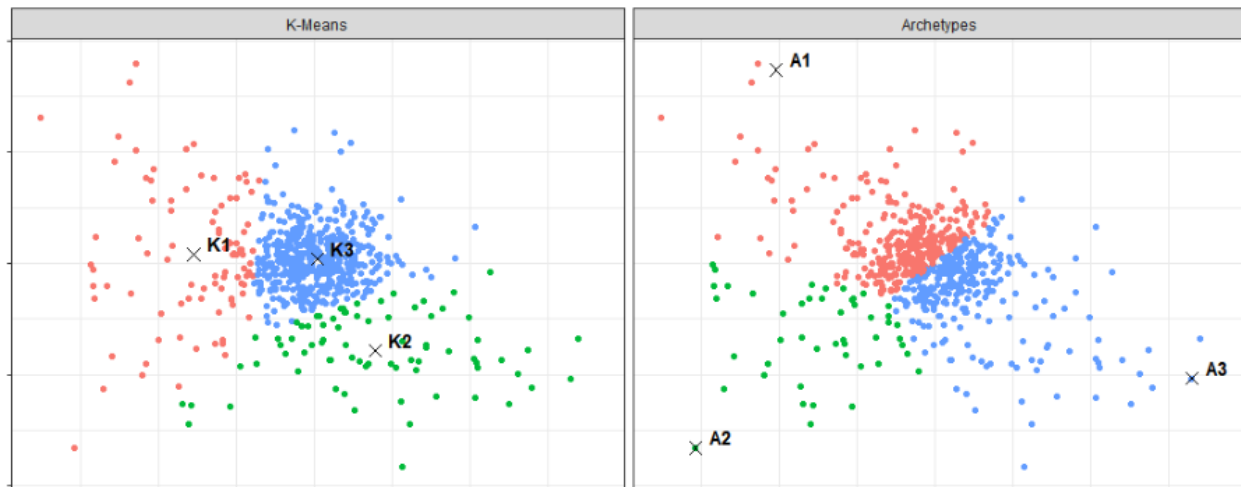
Conceptually, the centroid can be thought of as a prototype that embodies the characteristics of what it means to be a member of a cluster. For instance, the K1 centroid in Figure 1 represents a cluster of online shoppers (characterized by their online shopping behavior); it not only represents the average of that group, but also serves as a model for what it means to be a perfect member of that cluster. This prototype provides a benchmark against which other online shoppers in the data can be identified and assigned to the same cluster or segment.

IV. ARCHETYPAL ANALYSIS: LEVERAGING EXTREMES FOR MEANINGFUL SEGMENTATION

Much like K-Means analysis, archetypal analysis uses artificial data points to identify and exemplify the characteristics of data groups or market segments. However archetypal analysis stands apart from the traditional cluster analysis family. The artificial data points employed in archetypal analysis are known as “archetypes” and differ in their positioning within the data. While K-Means analysis centers on centroids representing average tendencies, archetypal analysis focuses on extreme observations located on the periphery of the data. Figure 2 illustrates a comparison of these artificial data points in a simulated dataset with two basis variables. The data points K1, K2, and K3 represent the centroids in K-Means analysis and A1, A2, and A3

represent the archetypes. Groups are assigned based on the proximity to the nearest centroid or archetype.

Figure 1



Using peripheral points to organize data into groups can significantly influence the outcome of segmentation. In cluster analysis, such as K-Means, the algorithm aims to identify homogenous groups where observations are highly similar to each other within the group. In contrast, archetypal analysis organizes groups based on contrastive categories defined by extremes, potentially resulting in *non-homogenous groups*.

Suppose the data depicted in Figure 2 were used for market segmentation. K-Means analysis identifies a very concise, homogenous cluster group represented by centroid K3, along with 2 diffuse cluster groups represented by centroids K1 and K2. Do segments defined by this analysis give researchers leverage for their market strategy? Perhaps not. The cluster defined by K3 exhibits no distinct characteristics in terms of the basis variables, except for being exceptionally average. While this group certainly exists within the market audience, it lacks usefulness for segmentation purposes. Despite fitting the data well, these groups offer limited leverage over market strategy. This phenomenon frequently occurs in market research, particularly when using multi-point Likert or semantic differential survey scales as basis variables. These scales often result in the emergence of a group of respondents that lean more towards the average on all variables.

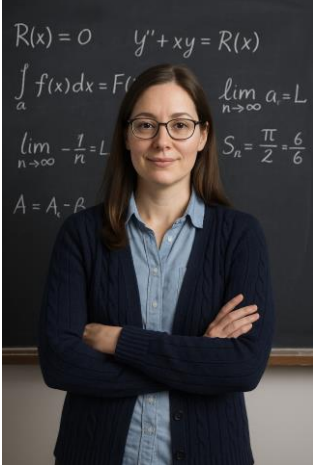
In such situations, it becomes more beneficial for researchers to categorize observations or respondents into groups defined by their tendency, whether strong or weak, towards data extremes. This approach is demonstrated on the right side of Figure 2, where the homogenous “average” group in the middle is split based on archetypes positioned on the data periphery, labeled as A1, A2, and A3. The middle group is mostly divided between A1 and A2. Contrasting with K-Means analysis, segments defined by archetypal analysis provide researchers with increased leverage over their market strategy. By defining segments in terms of sharp, contrastive categories rather than homogenous groups, researchers gain more actionable insights.

V. THE INTUITIVE NATURE OF ARCHETYPES

It’s important to recognize that categorizing observations using archetypes, or extremes, is a natural and intuitive process found in both nature and human psychology. This inherent intuition

makes the groups identified through archetypal analysis highly comprehensible and easy to understand. To illustrate this point, let’s consider two human categories: athletic and smart. When we think about the best examples to represent these categories, we tend to imagine individuals who embody the purest essence of athleticism or intelligence, rather than everyday instances. For example, the archetypes of athletic excellence might bring to mind figures like Michael Jordan, Serena Williams, or Michael Phelps. Similarly, for the smart category, archetypes such as Marie Curie, Albert Einstein, or Jane Goodall come to mind.

Examples of Archetypes

Athletic	Smart
	
	

Images generated by AI

In this context, an archetype can be thought of as the “champion” of a category. It represents the purest form or the highest embodiment within a class, serving as the standard against which all other members of the class are evaluated. Other individuals can belong to the same category if they possess qualities that at least partially resemble those of the archetype. For instance, a neighbor who jogs in the morning can be considered athletic because their qualities begin to approach those of the athletic archetype. Archetypal analysis aims to incorporate this intuitive psychological process into algorithms and market segmentation. For instance, in the context of market research, a market segment characterized by attitudes toward the environment can be identified by finding individuals within the market audience who hold extremely strong

environmental attitudes and using them as a benchmark to identify others with weaker but similar attitudes.

By leveraging archetypes, archetypal analysis provides a framework that aligns with our intuitive understanding of categorization, making it a powerful tool for segmenting and understanding complex market audiences.

VI. CONTRASTIVE CATEGORIES VS. HOMOGENOUS CLUSTERS

Archetypal analysis differentiates itself from cluster analysis by not requiring group or segment homogeneity, providing researchers with a unique advantage. Many attitudes and behaviors commonly used as basis variables in market segmentation exist on a continuous spectrum and lack homogeneity and conciseness, and so don't lend themselves well to traditional techniques.

An illustrative example is price sensitivity. Let's assume the researcher aims to identify a predominantly "price-sensitive" segment within a specific market audience, and a series of basis variables have been measured to assess this concept directly or indirectly. The survey data contains respondents with varying degrees of price sensitivity—some highly price-sensitive, some moderately price-sensitive, and some mildly price-sensitive. In other words, the condition of being price-sensitive lacks homogeneity. A cluster analysis, which seeks to find homogeneous groups, may not recognize these individuals as part of the same cluster or segment based on their price sensitivity. Instead, it may group them based on other variables or create separate clusters for "extreme," "moderate," and "mild" price sensitivity.

Contrarily, the researcher would prefer to have a single segment comprising price-sensitive individuals, regardless of the degree of their price sensitivity. Archetypal analysis, being distinct from cluster analysis, focuses on identifying the extremes that best summarize the data. It seeks to find sharp and contrastive categories to define groups or segments, sometimes at the expense of homogeneity within groups. In the case of price sensitivity, archetypal analysis would likely identify a group or segment of price-sensitive individuals within the market audience, despite the continuum of price-sensitive behavior. Many other common behaviors and attitudes used in segmentation also exist on a continuous spectrum and may benefit from the philosophy of archetypal analysis in market segmentation. Examples include luxury spending behavior, political attitudes, satisfaction levels, and more.

It's important to note that segment homogeneity is often valuable in market segmentation as it facilitates the creation of accessible and stable segments, thereby increasing their leverage over market strategy. Cluster analysis remains an indispensable tool in market segmentation research for this reason. However, segment homogeneity is not an absolute requirement in market segmentation. Enforcing homogeneity can hinder the discovery of powerful and contrastive segments, particularly when the measured basis variables do not naturally cluster in the data due to their continuous nature. In fact, researchers often prioritize understanding how segments differ from one another rather than how they are similar within groups. Archetypal analysis focuses on these differences between groups or segments, often leading to the discovery of more actionable segments compared to traditional cluster analysis.

VII. UNDERSTANDING THE MECHANICS

Gaining a solid understanding of how archetypal analysis works will help researchers effectively utilize this method and communicate segmentation results to others. While there is an

extensive body of literature delving into the theory, math, and machine learning aspects of archetypal analysis, a simple and intuitive explanation is all that is required to get started.

To illustrate the mechanics of archetypal analysis, let's consider a hypothetical scenario where a researcher collects data from 16 survey respondents regarding their attitudes toward online shopping behavior and luxury shopping behavior. Plotting this data on a simple XY scatterplot, we can envision stretching a rubber band around the outermost points of the data. The resulting shape formed by the rubber band is known as a “convex hull,” as shown in Figure 3 (Mørup and Hansen, 2012).

Figure 3

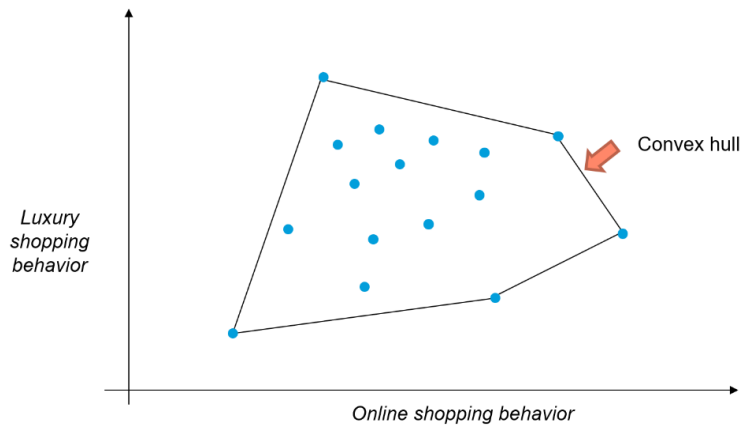
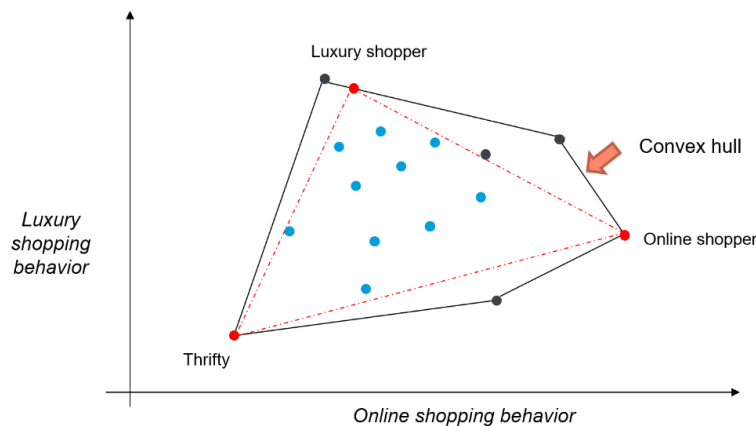


Figure 4



In Figure 3, we observe that the convex hull is defined by five endpoints. If we tasked archetypal analysis with finding five archetypes, the algorithm would converge on these endpoints as the archetypes. However, in practical market segmentation scenarios, researchers typically work with a larger number of basis variables and respondents, making the number of endpoints excessively large. Therefore, archetypal analysis aims to approximate the convex hull using a smaller number of archetypes.

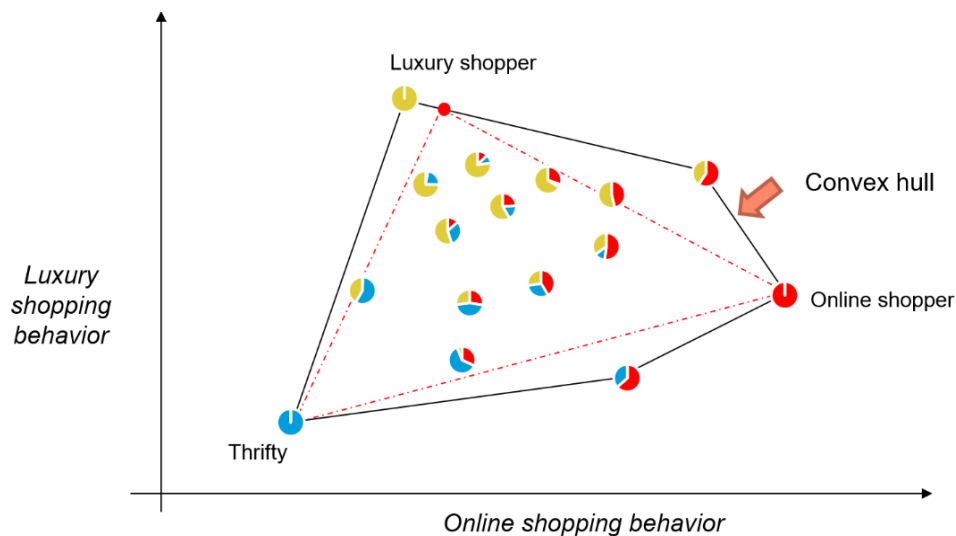
For instance, let's say the researcher wanted to find three archetypes in the example data. Archetypal analysis will then seek the three best endpoints or archetypes that “envelope” the

majority of the data. For the observations where it cannot, the algorithm strives to fit the data as close as it can minimizing the residual error of the model (the distance between the outside edge of the shape and the outside observations). The objective of the archetypal analysis is to minimize residual error as much as possible. Figure 4 illustrates how this shape is drawn.

With three archetypes identified, we can proceed to label them as market segments based on their values for luxury and online shopping behavior. In this example, the segments could be labeled as “Thrifty,” “Luxury Shopper,” and “Online Shopper.” Notably, these archetypes may not correspond directly to the original data points. It’s also worth mentioning that while the archetypes always exist along the convex hull, they may not necessarily be located on the endpoints.

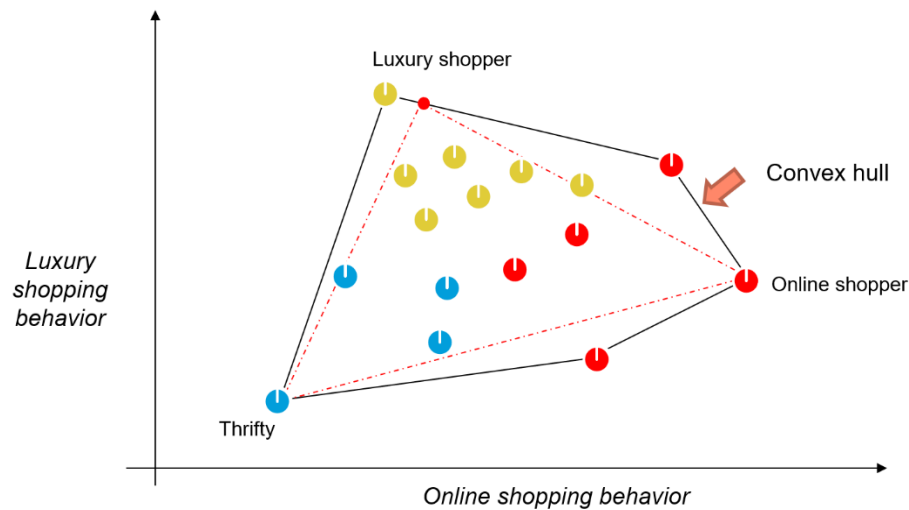
While fitting an archetypal analysis model, the algorithm generates what is known as the “alpha coefficient” for each respondent. These coefficients represent the relationship between each respondent and the archetypes. For every respondent in the data there is an alpha coefficient associated with each archetype. coefficients are constrained to be non-negative and sum to one for each respondent. As such, practitioners can treat them somewhat like propensity scores toward each archetype, i.e., each respondent has a proportion score assigned to them that represents their propensity toward each of the archetypes. Since these coefficients act like proportions, we can represent this graphically as mini-pie charts, as shown in Figure 5.

Figure 5



While the alpha coefficients hold intrinsic value, researchers will usually make more use of them by applying decision rules to assign respondents to specific segments based on their alpha scores. To do so, we should simply assign them to the segment or archetype where their alpha coefficient is highest. This is illustrated in Figure 6. Once all respondents are assigned a segment, the data can be treated like any other segmentation analysis, enabling the profiling of segments using additional variables. Furthermore, reporting additional statistics on the archetypes themselves becomes feasible.

Figure 6



VI. RUNNING ARCHETYPAL ANALYSIS FOR SEGMENTATION

Now that we have a basic understanding of how archetypal analysis works, let's explore how researchers can begin conducting their own archetypal analysis for segmentation projects. To perform archetypal analysis, the researcher will need to use a statistical programming language such as R. Fortunately, the process involves straightforward scripting, and there is a well-developed package called “archetypes” available on the CRAN repository in R that facilitates the analysis. This package is not only easy to use but also offers flexibility for advanced applications of the algorithm.

Installation and Usage

To get started, make sure the “archetypes” package is installed. You can do this by running the following command in R:

```
install.packages("archetypes")
```

Once the package is installed, the principal function to use in R is `steparchetypes()`. It's important to note that you should use this function instead of more basic functions like `archetypes()` to ensure proper handling of multiple starting points and avoidance of local convergence. The `steparchetypes()` function provides this capability.

An example implementation of archetypal analysis using `steparchetypes()` is given below:

```
library(archetypes)
data(toy)
set.seed(1)
archetypal_steps <- stepArchetypes(
  data = toy,
  k = 3,
  nrep = 25
)
```

In the example above, the ``data`` argument should be replaced with your actual data, and `vars` should be replaced with the variables of interest for segmentation. The `nrep` argument specifies the number of iterations or repetitions for the archetypal analysis. While the example uses four repetitions, it is often recommended to include more iterations (e.g., 25–50) to ensure convergence to the best model.

Exploring Multiple Numbers of Archetypes

The `steparchetypes()` function also allows for testing multiple numbers of archetypes. To do this, you can pass a numeric vector to the `k` argument. For example, using `k = 3:5` would build archetypal models with 3, 4, and 5 archetypes. The results of such a run can also be passed on to `screepplot()` to visualize how residual error varies among the proposed solutions.

```
archetypal_steps2 <- stepArchetypes(
  data = toy,
  k = 3:5,
  nrep = 25
)
screepplot(archetypal_steps2)
```

Extracting Alpha Coefficients and Assigning Segments

Once the model is run, you can use the `bestModel()` function to identify the best-fit model by examining all repetitions of the archetypal models. From there, you can extract the alpha coefficients, which represent the relationships between each observation and the archetypes. To assign segments to each observation based on the alpha coefficients, you can use the `max.col()` function.

```
best_model <- bestModel(result)
alpha_coefficients <- best_model$alpha
segments <- max.col(alpha_coefficients)
```

Further Resources

For more detailed explanations, tutorials, and documentation on the capabilities of the “archetypes” package, you can find online resources written by Eugster and Leish. These resources provide comprehensive guidance and can be accessed by entering `vignette("archetypes")` in the R console.

A full, more detailed explanation, tutorial, and documentation of the capabilities of the package can be found online, written by Eugster and Leish. They can also be found in R by entering `vignette("archetypes")` in the console (Eugster and Leisch, 2009).

VII. HANDLING OUTLIERS IN ARCHETYPAL ANALYSIS

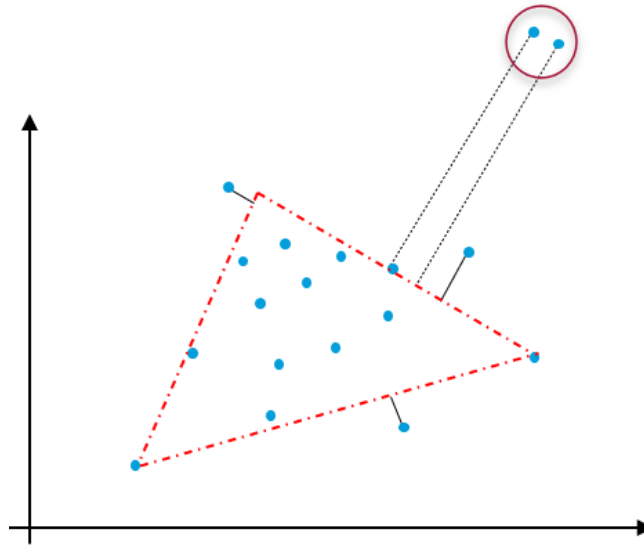
A common concern among researchers is whether archetypal analysis favors outliers in the data. While it is true that archetypal analysis privileges outliers, it doesn't necessarily pose a problem in all cases. The impact of outliers depends on their nature and how they influence the narrative of the model. In fact, extreme values are not inherently problematic for archetypal analysis and can sometimes reveal valuable and interesting segments in the data. The key is to run the model and evaluate the results to determine whether extreme outliers are unduly influencing the model.

Outliers in the data can sometimes be caused by measurement error. In such cases, researchers should handle them by applying data cleaning techniques as they normally would. It becomes particularly important to clean these outliers from the model when using archetypal analysis because the algorithm models based on the data periphery. However, there are situations where it is unclear whether an outlier is caused by measurement error or if it is accurately measured but exerting undue influence on the overall model. This can result in unhelpful archetypes, poor model fit, or both. In these cases, excluding a respondent's data from the model may not be desirable. Instead, researchers prefer the model simply give less or no attention to these observations.

The "archetypes" R package described earlier provides two modifications of the typical archetypes function that address the issue of outliers:

1. **Robust Archetypes:** This approach introduces a weight term to the residual error and solves for this term in the algorithm. The "archetypes" package offers this robust archetypes functionality, allowing the model to be less influenced by outliers when the multivariate distribution of basis variables significantly deviates from a Gaussian (bell curve) distribution. This is illustrated in Figure 8.
2. **Weighted Archetypes:** With this method, the researcher can assign custom weights to individual observations (respondents) and basis variables in the data. The "archetypes" package provides the weighted archetypes function, enabling the researcher to supervise the model by applying specific weights to control the influence of observations and variables.

Figure 8



In general, it is advisable to use robust archetypes when the multivariate distribution of basis variables deviates substantially from a Gaussian distribution. This helps account for non-normal data distributions and mitigates the impact of outliers on the model. On the other hand, if the researcher wishes to have more control and lightly supervise the model, the weighted archetypes function allows for the assignment of custom weights to observations and basis variables. (Eugster and Leisch, 2011)

VIII. CONSISTENCY OF ARCHETYPAL ANALYSIS

Consistency in segmentation refers to the stability and reliability of the segmentation algorithm's results when the analysis is repeated on different samples or subsets of the data. It measures the extent to which the clusters or segments identified in one dataset are consistent and reproducible in another sample or under different conditions. Assessing the consistency of segmentation results is crucial as it provides a measure of confidence in the identified solution.

A highly consistent segmentation solution indicates that the segments represent meaningful and stable patterns in the data, rather than being a result of random variation or specific characteristics of the analyzed sample. Understanding the consistency of a segmentation solution is valuable for confidently making data-driven decisions, deriving actionable insights, and formulating effective market strategies based on the identified segments.

Until recently, it was uncertain whether archetypal analysis produced consistent archetypes. However, in 2021, a team of mathematicians published a paper demonstrating the consistency of archetypes in archetypal analysis when the data was simulated with normal multivariate data distributions (Osting et al., 2021). This finding is encouraging. However, it is important to note that market research data is often not normally distributed, and researchers should consider this before assuming consistency in their specific segmentation solution. Furthermore, while the archetypes themselves may exhibit consistency, the decision boundaries used to assign respondents into segments might still lack consistency. More research is needed to evaluate the overall consistency of archetypal analysis as a segmentation tool.

In my career experience, I have observed that archetypal analysis generally exhibits stability in many situations, although there are exceptions. Challenges tend to arise when dealing with a high number of segments to identify, a low sample size, weak archetypes, and/or problematic outliers/basis variables in the data (refer to section VII). This is expected since such scenarios are more likely to have decision boundaries between segments that are closely located among observations. Consequently, the distribution of alpha coefficients tends to be flatter in these cases. To mitigate consistency issues, it is important to carefully select the appropriate basis variables and determine the number of archetypes based on the specific data at hand. Evaluating the distribution of alpha coefficients, the residual sum of squares error, and the consistency of segments through repeated resampling and analyzing the stability of the resulting segments can greatly assist researchers in avoiding consistency problems. Additionally, using a typing function can assist researchers in recovering the archetypal analysis segments in future datasets.

X. OTHER CONSIDERATIONS

Before concluding, we'll review two other aspects of archetypal analysis that the researcher should bear in mind, that of high dimensional data and using categorical variables and mixed data types.

High Dimensional Data

While high-dimensional data presents challenges in various segmentation algorithms, archetypal analysis offers certain advantages in this context. The representation of segments as extreme points allows archetypal analysis to capture the essential characteristics of the data, irrespective of its high dimensionality. The algorithm assigns weights to the basis variables based on their relevance in identifying the archetypes, effectively performing a form of pseudo-feature selection within the algorithm itself. Furthermore, archetypal analysis is non-parametric and does not heavily rely on assumptions about the underlying data distribution, enabling it to provide a flexible representation of the data structure. This flexibility proves beneficial in high-dimensional settings where meeting strict data distribution assumptions can be challenging.

Categorical Variables and Mixed Data Types

Like many other commonly employed segmentation methods, archetypal analysis is primarily designed for numerical data. However, this creates a dilemma for researchers dealing with survey data that includes a mix of continuous, ordinal, and categorical variables, requiring the use of different variable types in segmentation. While using mixed data types can be problematic in other segmentation algorithms, the effect on archetypal analysis differs. For instance, if a categorical variable is “dummy-coded” into a series of 1s and 0s and included in archetypal analysis, these binary values would always reside along the convex hull of the data, influencing the generation of archetypes more than the numerical variables. Handling categorical or mixed data types in archetypal analysis is challenging, but one approach is to apply custom weights to the basis variables to even out their influence (refer to section VII).

XI. CONCLUSION

By incorporating archetypal analysis into the practitioner's toolkit, researchers are empowered to unlock new types of segments beyond what traditional cluster analysis techniques offer. The utilization of extreme observations, embrace of contrastive categories, and

consideration of variables on a long continuum make archetypal analysis a compelling approach for creating actionable, intuitive, and narratively rich market segments. While it may not always be the optimal choice for every segmentation algorithm or occasion, archetypal analysis remains an important tool in every practitioner's arsenal. With its potential to align with research objectives and deliver valuable outcomes, archetypal analysis stands as a valuable asset for practitioners seeking to elevate their market segmentation practices.



Jacob Nelson

CITATIONS

- Chrzan, K. (2021, February 15). *Segmentation: Four common types of segmentation (part 1)*. Survey Software & Market Research Solutions—Sawtooth Software. <https://sawtoothsoftware.com/resources/blog/posts/four-common-types-of-segmentation>
- Cutler, A., and Breiman, L. (1994). Archetypal analysis. *Technometrics*, 36(4), 338–347. <https://doi.org/10.1080/00401706.1994.10485840>
- Eugster, M. J. A., and Leisch, F. (2011). Weighted and robust archetypal analysis. *Computational Statistics & Data Analysis*, 55(3), 1215–1225. <https://doi.org/10.1016/j.csda.2010.10.017>
- Eugster, M. J., and Leisch, F. (2009). From spider-man to hero—archetypal analysis in *r*. *Journal of Statistical Software*, 30(8). <https://doi.org/10.18637/jss.v030.i08>
- Li, S., Wang, P. Z., Louviere, J. J., and Carson, R. (2003). Archetypal analysis: A new way to segment markets based on extreme individuals. In Australian and New Zealand Marketing Academy Conference. ANZMAC.
- Mørup, M., and Hansen, L. K. (2012). Archetypal analysis for machine learning and Data Mining. *Neurocomputing*, 80, 54–63. <https://doi.org/10.1016/j.neucom.2011.06.033>
- Osting, B., Wang, D., Xu, Y., and Zosso, D. (2021). Consistency of archetypal analysis. *SIAM Journal on Mathematics of Data Science*, 3(1), 1–30. <https://doi.org/10.1137/20m1331792>

INTEGRATING CONSUMER GOALS IN CONJOINT USING ARCHETYPES

MARCO VRIENS

KWANTUM

DARIN MILLS

ANDREW ELDER

ILLUMINAS

SUMMARY

In situations where the market is heavily dominated by brand, price, and design it can be challenging to find product feature changes that are impactful enough to change the market share landscape. In markets where product features change very fast, conjoint results can be short-lived as whichever feature is most important and may be distinct today will be a commodity and not important tomorrow. We illustrate that by integrating goals and benefits into a conjoint analysis we can mitigate both situations. We show that by using Archetypal analysis we can identify switchable consumers. These are more prone to respond to changes in product features. We also illustrate that by integrating goals and benefits into our conjoint model, we get more strategic insights that have a longer shelf life.

INTRODUCTION

It is well known with conjoint analysis that the resulting insights into attribute level utilities in combination with a market simulator can be used to optimize which attribute combinations maximize expected market share. However, there are situations where such information is not fully sufficient to extract actionable insights. Two factors specifically can have a big impact on how useful the conjoint is and over what time. One, there are situations where the product choices are more heavily dominated by brand and price (and in our study, form factor, as opposed to other (micro) attributes). Two, in technology-driven markets, the set of available features can change quickly. A newly tested feature in the conjoint can be seen as a commodity or even obsolete because new features enter the market fast. In this paper we outline an approach that helps us get actionable strategic insights when these two factors are at play. We propose to integrate consumers' goals and perceived product benefits with the conjoint results using Archetypal analysis.

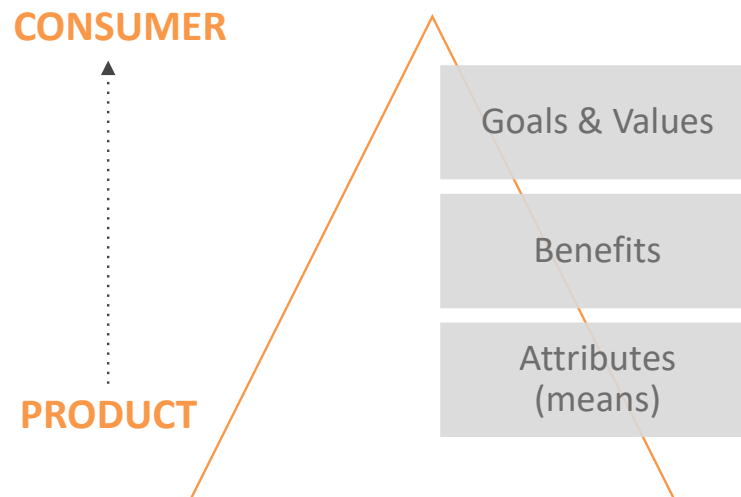
In the next section, we discuss the value of consumer goals and benefits. In section three, we outline the survey design and analysis steps. In section four we present some selected key results. Lastly in section five, we offer some key takeaways.

THE VALUE OF GOALS AND BENEFITS

Goals and values are foundational drivers of consumer behavior (Gutman, 1982; Van Osselaer and Janiszewski, 2011). Integrating goals with conjoint has two practical benefits. One, it helps with integrating product and marketing decisions. Two, by linking goals to attribute utilities we can extend the life span of the conjoint as goals are typically more stable than preferences for specific features.

A well-known framework that links the importance of product attributes to benefits and goals is the means-end chain framework. See Figure 1.

Figure 1. The Means-End Chain Framework



The framework shown in Figure 1 has been used in consumer research for decades. Attributes have value because they lead to benefits and/or goals, and they provide in essence the first reason why. Benefits have value because they are associated with achieving certain benefits or goals. It is a nice way to link the consumer to a product as attributes are completely features of products and goals are completely features of consumers. Goals are assumed to be enduring motivators of consumer choices. For example, thinking of fitness bands and wearables, step counting (a product attribute) can lead to losing weight (a benefit), and losing weight can be seen as an important aspect of improving overall health (goal). You can see how understanding the importance of the goal can extend the longevity of conjoint results, because even as step counting by now has become a commodity feature, the consumer goal, improving health probably hasn't changed. If I know that losing weight is more important than less stress, then as new features become available that elicit these benefits, the feature associated with losing weight is probably more important.

Several methods have been proposed to integrate product attributes, benefits, and consumer goals. This overview below is probably not completely comprehensive, but it shows some existing methods to connect attributes to goals. One of the oldest methods is probably the laddering method (e.g., Vriens and ter Hofstede, 2001), where we literally ask people what attributes they see connected with which benefits and goals and values, and which benefits they see connected with which goals and values. The downside of this method is that we don't get attribute level utility values, although laddering can be combined with conjoint. Another method is benefit conjoint (see Kim et al., 2017). This model is similar to the model proposed by Wedel et al. (1998). The problem with this approach is that we don't really know what benefit it is that consumers are seeing. All we know is that a certain combination of attributes shares a latent variable. If these attributes have a certain theme in common, then we may interpret that as a benefit, but it is not guaranteed that this will happen. A third method explicitly links attributes and benefits in different conjoint designs and is referred to as Hierarchical conjoint (e.g., Oppewal, Louviere and Timmermans, 1993; and Oppewal and Vriens, 1998). This method is a little convoluted, and not super practical we think. Lastly, we can try to ask about benefits and

goals using Archetypal analysis (e.g., Liu, Korz and Allenby, 2023). Our approach is similar. It has the advantage that it is transparent and easy to implement.

SURVEY DESIGN AND ANALYSIS STEPS

In this section, we briefly outline the survey design that we used to capture the attribute tradeoffs, benefits and consumer goals and we outline the analysis steps.

Survey Design

There are 3 components in the survey that involve our methodology: Benefits, goals, and a conjoint exercise. Specifically, we included: 12 benefits, 15 goals and we used 2 conjoint exercises: a Macro and a Micro conjoint. These are the elements we are trying to tie together in our analysis.

The benefits section in our study was presented in a series of semantic differential tradeoffs. For example:

For me brand is very important.....I believe most brands are more or less the same.

Respondents then subsequently indicate on a 4-point scale whether the left statements describe them more or whether the right statement describes them more. As an additional benefit, such semantic differential benefit exercises have been tested in conjoint studies and can improve the conjoint responses (e.g., Kurz and Binner, 2021).

The goals were simply listed, and respondents had to indicate whether the stated goal applied to them; simple Yes/No questions. They ranged from general health, general fitness, to more specific health or fitness goals. For example, goals such as “Reduce stress,” “Live pain free,” “Reduce my A1C,” and “Get Stronger.”

We have two conjoint exercises, a macro, and a micro design. The market in our case has a wide price range, from as low as \$ 50 to as high as \$ 700. Also, the brands in this market were distinct and choosing a particular brand could affect other products the consumer was using. Last, the full list of attributes was large. Hence, the conjoint was structured as follows:

1. A Macro Conjoint

The Macro conjoint only included brand, form factor, and price as the attributes along with product images so respondents could better identify products. Actual product combinations were used in half the tasks to ensure current market tradeoff choices. The specific prices shown for each product were rotated but only within +/- 1 level of the actual product price. This was done because the main purpose of the Macro conjoint was to channel respondents into the appropriate price range of their perceived preference while still being able to inform some price elasticity. Respondents saw eight choice sets, each set containing six alternatives (including a none option).

Based on their selections in the Macro conjoint, each respondent was allocated to a Low/Mid/High price band. As shown below in Table 1, we allowed there to be an overlap in the prices shown between each price band. Knowing that the attributes tested in the Macro conjoint can heavily influence the product choice, the brand shown was not constrained to any specific price band.

2. Micro Conjoint

The Micro conjoint was utilized to gain insight into the value of various health, fitness, and safety features. Respondents in different price bands would get exposed to different price levels. The Micro conjoint was set up so that there was some overlap between the low and mid-range price bands and some overlap between the mid and high price range. See Table 1 below.

Table 1: Price Levels Across the Three Price Bands

Price Shown in Micro	Price Band		
	Low	Mid	High
Price 1	X		
Price 2	X		
Price 3	X	X	
Price 4	X	X	
Price 5	X	X	X
Price 6		X	X
Price 7		X	X
Price 8			X
Price 9			X

The Micro conjoint was designed with 16 attributes. Respondents saw 11 sets, each set containing four alternatives including the none-option.

In addition to the brand, form factor, and price attributes, some examples of the Micro conjoint attributes we tested were:

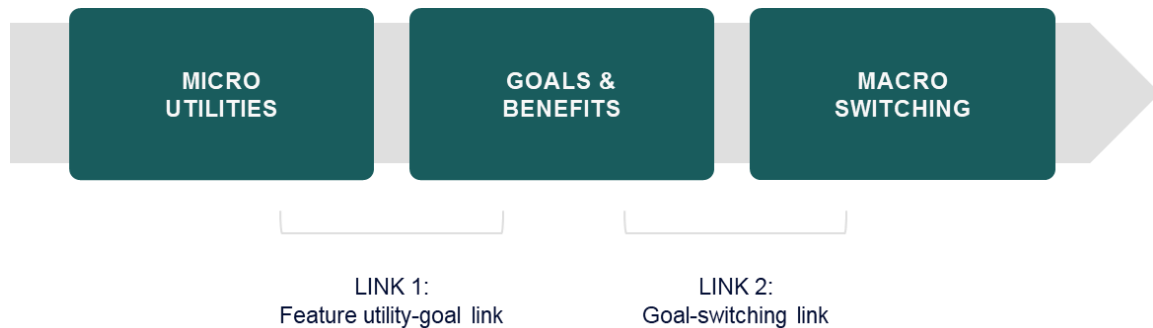
- Safety
- Stress
- Training
- Tracking
- General Health
- Sleep

Analysis Steps

While we now have many tasks with many attributes to model, the first step was to determine the best modeling approach to combine the Macro and Micro conjoint. Ultimately, we chose HB utility estimation using tasks from each conjoint to optimize the mean absolute error (MAE) and hit rate.

As we expected, brand, price and form factor were dominating attributes in the model. This limited our ability to analyze which specific health/product attributes led to brand switching. Or to put another way, which features (or combination of features) enticed the largest number of respondents to switch brands. So, we decided to pull in the benefits and goals. The analysis framework is shown in Figure 2.

Figure 2: Analysis Framework



To home in on respondents most likely to switch brands we proceeded as follows:

Step 1. Determining Who is Open to Brand Switching

In step one we aim to identify those respondents most open to switching between brands. For this we used Archetypal analysis (e.g., Cutler and Breiman, 1994) on the macro conjoint data. Preferences were strongly brand driven and hence we expected to find archetypes around preference for the tested brands A, B and C.

Archetypal analyses allow us to look at each respondent's probability of belonging to an Archetype. This benefit allows us to identify those respondents who have similar probabilities across multiple segments without having a dominant archetype. In our case, we decided a respondent has a dominant brand archetype if their probability of association (coefficient) is more than 0.5. As we detail in the results section, this allows us to differentiate between respondents who are not likely to be open to switching between brands (i.e., Brand Loyalists) and those who are open (i.e., Brand Switchers). This "Switchable Consumer" designation becomes our dependent variable later.

Archetypal analysis is only one of many partitioning methods available to help identify Brand Loyalists. Unlike most consensus or distance-based methods (e.g., k-means clustering, ensemble clustering, etc.; see Vidden, Vriens and Chen, 2016), the archetypal coefficients clearly articulate switching opportunities, rather than simply identifying areas of uncertainty between classification. This structure had the additional benefit of representing market share across competitors better than other partitioning solutions. In this study, the impact of a dominant brand exists beyond just those who are loyalists. Archetypal analysis does a better job than other approaches at revealing the subtle impact of a dominant brand, even among those individuals who have mixed brand preferences.

Step 2. Incorporating Goals and Benefits

The next step in our analysis is an archetypal analysis of the goals and benefits. We expected respondents to differ regarding the number of goals, and we expected that respondents would be different in the types of goals (e.g., some more health focused, others fitness focused).

In the next steps, we use both the goals and benefits data directly and we use the archetypal goal segments and archetypal benefit segments.

Step 3. Predicting Brand Switching Based on Goals and Benefits

In the third step, we ran Decision Trees (Breiman, et al., 1984) with the Switchable Consumer designation (yes/no) as the dependent variable. The independent variables are 1) the goal archetypes membership probabilities, 2) the goals variables directly, 3) the benefit archetypes probabilities, and 4) the benefits directly.

Step 4. Linking Micro Conjoint Utilities to Goals

Once we derived the connection between the Switchable Consumer and the goals/benefits, we tied the specific features tested in the Micro conjoint to the goals and benefits using regression analysis.

RESULTS

First, we looked at the macro conjoint choices knowing that brand, price, and form factor can heavily dominate in the product choice process. We used Archetypal analysis on the stated macro choices. Then, we use the probabilities of association to identify those with high or low brand preferences. Look at Table 2 below.

Table 2: Example of Coefficients from Brand Preference Archetypal Analysis

	Probability of Association (Alpha Coefficient)			
Respondent	Brand A	Brand B	Brand C	
1	0.1	0.7	0.2	
2	0.3	0.2	0.5	
3	0	0	1	--> Strong Brand Affinity
4	0.2	0.2	0.6	
5	0.2	0	0.8	
6	0.3	0.35	0.35	-->Uniform Brand Affinity
7	0.6	0.2	0.2	
8	0	0.9	0.1	--> Strong Brand Affinity

As Table 2 shows, there are some respondents with a dominant probability for one archetype (for example, respondents 3 and 8). However, there are also some respondents whose probabilities are very similar across two or three brands (like respondent 6 in Table 2). In essence, we are identifying those who do not have a strong brand affinity and are more likely to switch brands. We have dubbed these as “switchable consumers,” then used this switchable consumer designation as the basis for understanding brand preference.

Next, we derived two archetypal solutions. One for the binary health goals and another for the semantic differential product benefits statements. Archetypal analysis of benefits yielded four archetypes (not shown in this paper). Below, we are only showing the profiling for the health goal archetypes. We have incorporated both solutions in the rest of the analysis. See Table 3 below.

Table 3: Goals Archetypes

Health Goal Archetypes	A1	A2	A3	A4	A5
Descriptive Labels	Unmotivated	Maintain Health	Fitness Improvers	Become Healthy	Better in everything
Specific Goals	Few	Some	Several	Many	Exhaustive
Top Goals	Goal1	Goal1; Goal2	Goal1; Goal3; Goal4	Goal1; Goal2; Goal4; Goal5	Goal1; Goal2; Goal3; Goal4; Goal5; Goal6; Goal7

After generating the two archetype solutions on health goals on product benefits, as well as identifying the switchable consumers designation, we wanted to find out if we could predict whether a respondent was a switchable consumer using the goals and benefit data. Several statistical methods can be used for this, but we settled on using a decision trees (DT) analysis (Breiman et al., 1984). One of the benefits of DT is that it automatically identifies interaction effects.

The results of the first DT analysis are shown below in Figure 3.

Figure 3: Decision Tree with Goals and Benefit and Macro Archetypes as Independent Variables

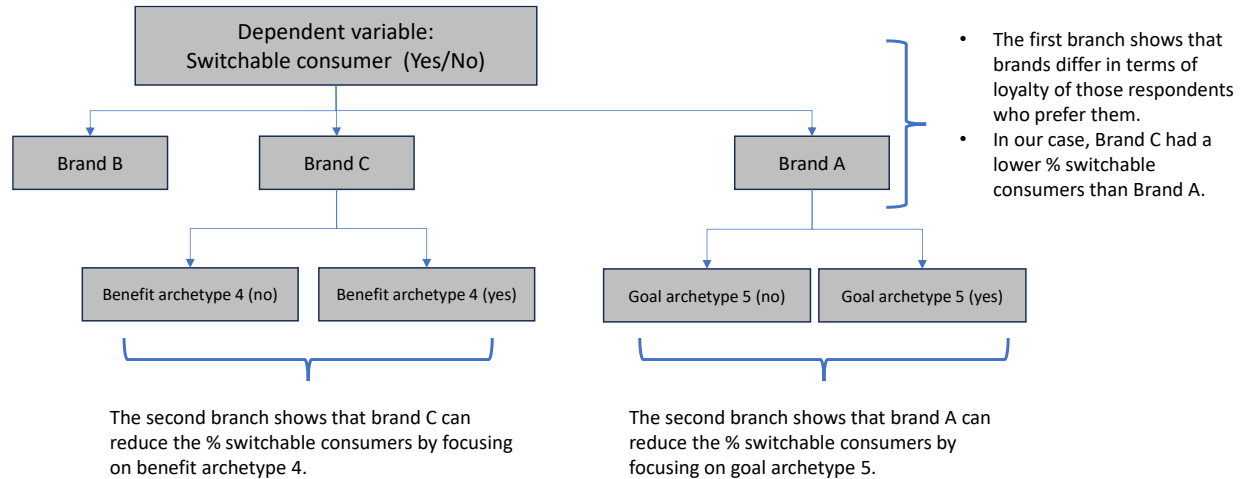
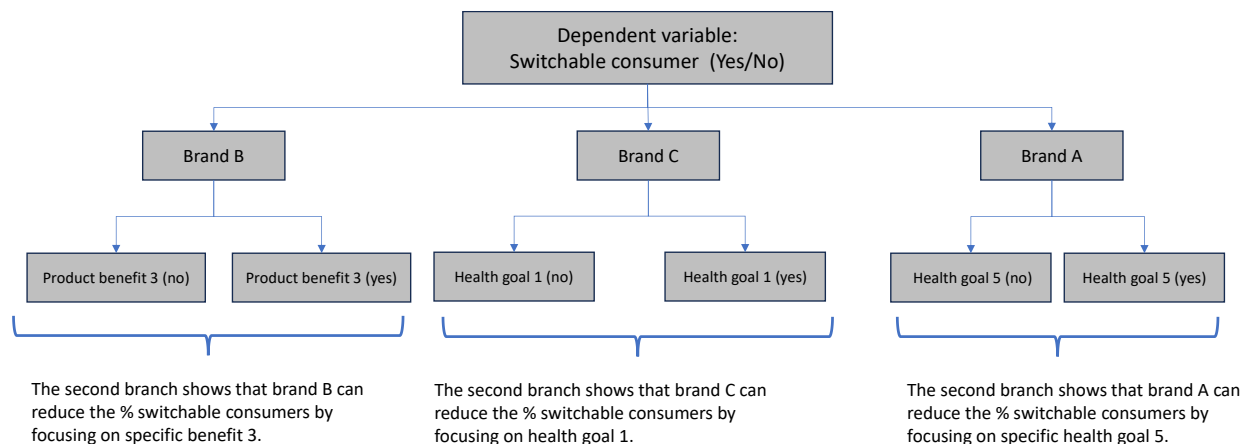


Figure 3 is a simplified representation of the actual decision tree. We don't show the actual differences in percentage switchable consumers in each brand. In branch 1, brand C had a lower number of switchable consumers than brand A. In branch 2, under brand C, the biggest difference in percentage of switchable consumers was found between archetype 1 and 4, where respondents who score high on archetype 4 had the lowest percentage switchable consumers. In branch 2 under brand A, the lowest percentage switchable consumers were found for those scoring high on goals archetype 5.

It shows two key insights. First, the first branch shows that the three brands in our study differ with respect to the percentage of switchable consumers. This was not entirely surprising, but it was still useful to see the magnitude of this "brand loyalty." Second, we can see that brand C needs to compete on benefits whereas brand A needs to leverage how consumers view it as instrumental in achieving certain goals. This is an important strategic insight for the product roadmap process. Next, is there a way to identify or predict who is more likely to switch (to switch from your brand, or to switch from a competitor's brand)?

What we are missing in this decision tree is what predicts brand loyalty for brand B. To investigate this further, we ran the DT using the individual benefits and goals that make up each archetypal solution to see if those did a better job of teasing out these differences (see Figure 4 below).

Figure 4: Decision Tree with Specific Goals and Benefit and Macro Archetypes as Independent Variables



Note: The decision tree shown above is a simplified version. The actual tree had multiple branches. For the sake of simplicity, we are only showing two branches. There are two key insights here. One, in this tree, we do find what differentiates respondents who are loyal vs. less loyal for brand B. Further branches of the tree showed that loyalty for brand B hinges mostly on whether the respondents require very specific product benefits. Two, this decision tree gives us more tactical insights as it identifies very specific benefits and goals.

The next step was to tie the health and benefit archetypes to the specific attributes tested in the micro conjoint. If we know the goals can help predict and add context to brand switching, what specific health attributes best predict the health goals: i.e., link 1 in Figure 2. To answer this, we modeled the number of health goals as a function of the attribute utilities. See Table 4 below.

Table 4: Regression Results (*disguised*)

Independent variables		Coefficient	p-value
Constant		1.73	0
Safety		0.01	0.31
Stress		0.04	0.49
Training		0.47	0
Tracking		0.22	0
General health		0.01	0.77
Sleep		0.22	0

Note: adjusted $R^2 = 0.36$

These regression results exposed the features that best forecast the health goals and product benefits separately across all brand archetypes.

If we want to complete the linkage from attribute features to goals to switching, we need to subset the sample by each of the brand archetypes to have a concrete roadmap for each brand. So, referring to the Decision Tree where brand A switching revolved around health goals (Figure 3), we can tie the filtered (brand A archetype) regression's significant contributors (denoted with ***) to the health goals and ultimately to the brand switching. Now we can connect the dots in tackling how to turn health goals into integrated product features. In other words, the attributes identified in the regression can be positioned to capture health goals that lead to switching.

Then brand A can use this high-level roadmap to influence product development and marketing outreach from a defensive position to retain likely switchers. Conversely, brand C could use those findings to attract likely brand A customers.

For this paper due to time constraints, we only included these findings tying the health goals with the health attribute utilities, but we repeated the analysis focusing on the benefit archetypes as a function of the utilities with similar findings.

CONCLUSIONS AND TAKEAWAYS

Incorporating goals and benefits have humanized the conjoint analysis and extended the longevity of the findings as they are foundational consumer elements that remain stable longer than preferences specific product features.

Second, in situations where consumers' choices are heavily dominated by brand, it is hard to extract specific attribute-level insights that can inform product decisions and the product roadmap. By identifying the switchable consumer, via an Archetypal analysis of the macro conjoint in our study, we were able to extract insights that are not visible at the overall sample level. This showed both a strategic insight into how differently the different brands should compete, and for our client yielded insight into an effective high level strategic product roadmap.



Marco Vriens



Darin Mills



Andrew Elder

REFERENCES

1. Breiman, L., Friedman, J., Olshen, R.A., and Stone, C.J. (1984). Classification and Regression Trees. CRC Press.
2. Cutler, A. & Breiman, L. (1994). Archetypal analysis. *Technometrics*, 36, 338–347.
3. D’Esposito, M.R., Palumbo, F. and Ragozini, G. (2006). Archetypal analysis for Interval data in marketing research. *Statistica Applicata*, 18, 2, 343–358.
4. Gutman.J. (1982), “A Means-End model based on consumer categorization processes, *Journal of Marketing*, 46, 60–72.
5. Kim, D.S., Bailey, R.A., Hardt, N., Allenby, G.A. (2017). Benefit-based conjoint analysis. *Marketing Science*, 36, 1, 54–69.
6. Kurz, P. and Binner, S. (2021). Enhance conjoint with a behavioral framework. In: *Sawtooth Software Conference Proceedings*, 91–107.
7. Liu, Y-C. M., Korz, P., and Allenby, G. (2022). Archetypal analysis and product line design. In *Sawtooth Software Conference Proceedings*, Orlando, FL., 255–267.
8. Oppewal, H., Louviere, J.J. and Timmermans, H. (1993). Modeling hierarchical conjoint processes with integrated choice experiments. *Journal of Marketing Research*, 31, 1, 92–105.
9. Oppewal, H. and M. Vriens (2000). Measuring perceived service quality using integrated conjoint experiments. *International Journal of Bank Marketing*, Vol. 18, No. 4, pp. 154–169.
10. Van Osselaer, S.M.J. and Janiszewski, C. (2011). A goal-based model of product evaluation and choice. *Journal of Consumer Research*, 39, 260–292.
11. Vidden, C., M. Vriens, S. Chen (2016). Comparing clustering methods for market segmentation: A simulation study. *Applied Marketing Analytics*, 2, 3, 225–238.
12. Vriens, M., Ter Hofstede, F. (2001). Linking attributes, benefits, and values: A powerful framework to market segmentation, brand positioning & advertising strategy development. *Marketing research magazine*. 3–8.
13. Wedel, M., M. Vriens, T. Bijmolt, W. Krijnen and P.S.H. Leeflang (1998). Assessing the effects of abstract attributes and brand familiarity in conjoint choice experiments. *International Journal of Research in Marketing*, Vol.15, pp. 71–78.

THE IMPACT OF MULTIPLE CLUSTER STRUCTURES ON VARIABLE SELECTION IN SEGMENTATION

JOSEPH WHITE
KYNETEC

ABSTRACT

Effective basis variable selection for cluster analysis addresses a set of challenges that can adversely affect the outcome of segmentation studies. Previous research shows that the R package *clustvars* performs well at identifying the correct basis variables for a single known cluster structure. This paper leverages an experimental design and synthetic data to show the impact of multiple complete cluster structures in our data on variable selection effectiveness. The R package *clustvars* outperforms a newly considered R package, *VarSelLCM*, and a manual selection technique based on random forests in terms of selecting effective variables, identifying the right number of segments, and accurately classifying records.

INTRODUCTION

Successful segmentation studies require analysts to balance researcher desires, client needs, and analytic rigor. The former two often lead to a plethora of basis variables to be used in defining clusters. In spirit this is a noble cause, we want to capture the nuances of heterogeneity between respondents, but in practice this leads to data challenges that can hamper our ability to find the true cluster structure when one exists.

Previous research shows that effective variable selection improves our chances of finding the true cluster structure (Chrzan and White 2022). However, that research only considered the case of a single structure in the data. Given the amount of data often available for segmentation, be that due to long surveys or rich customer databases, the existence of multiple cluster structures in our data is likely a common occurrence. The focus of this paper is then to understand the impact of multiple structures on effective variable selection and subsequently the segmentation quality, i.e., what is it that we get?

DATA CHALLENGES TO SEGMENTATION

There are many data challenges that analysts face in segmentation studies, some of which may be alleviated by effective variable selection. Some data issues that effective variable selection can help mitigate include the curse of dimensionality, sample size requirements, masking variables, and correlated measures, discussed briefly in turn.

The Curse of Dimensionality

An excellent description of this problem comes from Yiu (2019):

When we have too many features, observations become harder to cluster—believe it or not, too many dimensions causes every observation in your data set to appear equidistant from all the others. And because clustering uses a distance measure such as Euclidean distance to quantify the similarity between observations, this is a big problem. If the distances are all approximately equal, then all the observations appear equally alike (as well as equally different), and no meaningful clusters can be formed.

This is a phenomenon observed routinely when working with data from surveys with large numbers of attributes intended to serve as basis variables. It is easily seen in silhouette plots as you increase the number of basis variables in your clustering. The more you add the less clear the groupings as evidenced by the narrower silhouettes, meaning your data become harder to cluster into meaningfully distinct groups. Effective variable selection directly helps to avoid the curse of dimensionality.

Sample Size

As with most, if not all, multivariate analyses, the need for sample size increases as the number of variables increases. A few general rules of thumb for segmentation studies are:

Formann (1984)	$n \geq 5 * 2^d$
Qui and Joe (2009)	$n \geq 10 * d * k$
Dolnicar et al. (2016)	$n \geq 100 * d$

Where d is the number of basis variables and k the number of clusters. Given a modest space of 20 basis variables, and a typical solution of 5 clusters, these rules of thumb suggest 5,242,880, 1,000, and 2,000, respectively. Clearly, the Formann rule of thumb is infeasible for marketing researchers, and while the other two are sometimes attainable it is the author's experience that even these sample sizes are often out of reach given budget constraints and/or target populations.

Masking Variables

Masking variables (Brusco 2004) are variables that serve only to hide the latent cluster structure in your data. These variables have no relation to the cluster structure and serve as noise to segmentation algorithms. Inclusion of masking variables interferes with distance calculations and contributes to the problems of dimensionality noted above.

Last year at the Sawtooth Software Conference we found that identifying and removing masking variables was a simple task for the automated selection techniques tested as well as the manual ANOVA selection process (Chrzan and White 2022). Due to this and the impact of including too many variables on processing time, masking variables are not a focus here.

Correlated Variables

Including correlated measures of dimensions in your basis variables harms your ability to correctly identify a cluster structure (Chrzan and White 2021, Dolnicar et al. 2018). Practitioners are likely accustomed to working with survey data where multiple attributes are used to measure a hypothesized construct, especially in attitudinal questionnaires. Being able to effectively select the best among a set of correlated indicators then not only helps with dimensional challenges but also improves the likelihood that we recover the underlying structure.

VARIABLE SELECTION PROCEDURES

This research considers three variable selection techniques. All procedures are carried out in R, with one being a manual approach and two leveraging the R packages written for variable selection, *clustvarsel* (CVS) and *VarSelLCM* (VSL).

The CVS package was found to perform exceptionally well at effectively selecting a set of basis variables last year, so it is included again to see what happens when facing the more complex issue of competing cluster structures. The drawback to CVS is processing time, which averaged about 20 minutes per iteration in simulation.

The second R package, VSL, is significantly faster than CVS so is included as an alternative. It is also able to handle mixed data types, a bonus if the results are on par with or better than CVS.

The manual process starts with calculating a proximity matrix with an unsupervised random forest (RF) using the R package *randomForest*. The proximity matrix then supports clustering via partitioning around medoids (PAM) using the R package *cluster*. A stepwise discriminant analysis is then performed to identify a reduced set of variables for the final clustering.

Finally, all reduced sets of variables are clustered using *mclust*, a model-based clustering package in R.

HYPOTHETICAL EXAMPLE

The question at hand is how variable selection is impacted by coexistent cluster structures in our data. Before turning to the research design, a simple example helps to illuminate the issue.

Consider the hypothetical example of segmenting small business owners in your favorite industry. It would not be uncommon to conduct some qualitative research as a first step to understand potential dimensions, and to then come up with a list of attributes that describe owner attitudes. Suppose after going through this process we end up with four general attitudinal buckets, Risk, Technology, Business Partner Relationships, and Community Involvement. And that we identify five attributes for each.

It is entirely possible that there is a complete and well-behaved cluster structure based on attitudes towards Risk and Technology alone. At the same time there could be an entirely different cluster structure equally well defined, based on attitudes towards Business Partner Relationships and Community Involvement.

A segmentation based on just our Risk and Technology attributes should, after effective variable selection, do a good job of uncovering that true cluster structure. Similarly, clustering on only our Relationships and Community attributes should do a good job of uncovering that true segment structure.

What emerges when we apply our algorithms to the combined data is the purpose of this paper. Do we end up with the full joint structure, one or the other, or something in between?

RESEARCH OVERVIEW

The research strategy involves four key phases.



In the define phase the individual cluster structures are specified. The strategy is to first create well-defined single cluster structures that are later merged to establish coexistence. Each individual cluster structure has data challenges/characteristics that vary according to a factorial design.

Once the structures are specified, artificial data are generated accordingly in such a way as to ensure complete separation. This represents a best-case scenario.

The independently generated single-cluster data files then need to be combined in a manner conducive to analysis. Considerations of independence, single structure characteristics, and processing time come into play at this stage.

Finally, the combined data files are analyzed. The success at identifying the right number of clusters, effectively removing redundancies, and accurately assigning records to their true segment are explored.

DEFINING SINGLE STRUCTURE

Single cluster structures are defined according to a factorial design based on different characteristics that can adversely affect our ability to recover the true segments. Accordingly, each single structure is an individual data set with a combination of the following characteristics.

Factor	Level 1	Level 2
Number of Segments	3	5
Segment Size	Even	Uneven
Dimensions	3	5
Indicators per Dimension	1	5
Separation	Small	Large

All simulated data sets contain 685 records to facilitate merging. Uneven segment sizes are approximately $373/k$ for the k^{th} cluster when we have 3 segments, and $300/k$ for the 5-segment structure.

Thus, we have a 2^5 design space resulting in 32 different segment structures to simulate. Each cell is replicated 40 times giving a total of 1,280 independent structures from which to select for analysis.

DATA GENERATION

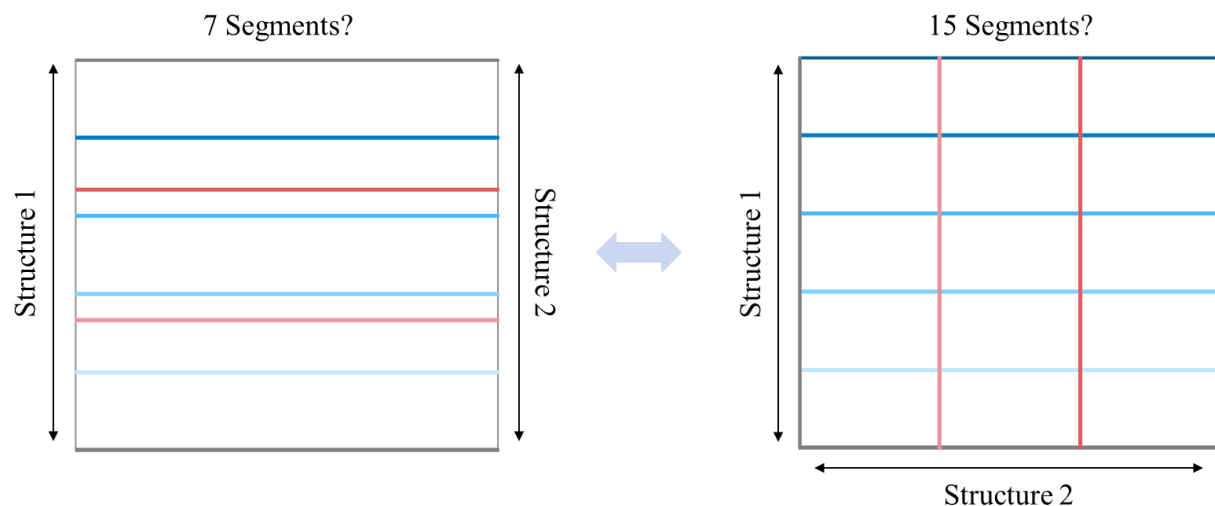
As previously mentioned, data are generated so there is complete separation between clusters within a single structure. The algorithm for simulating the data is as follows.

1. Draw initial centroids from a multivariate normal distribution with zero mean and variance according to the design cell.
 - a. Small separation: $\sigma^2 = 1$
 - b. Large separation: $\sigma^2 = 6$ (Qiu and Joe 2020)
2. Generate a pool of candidate records by taking random draws from a multivariate standard normal distribution with covariance of 0.8 between variables within a dimension and 0 otherwise.
3. Randomly assign to one of the segments.
4. Shift data according to the centroid location, standardize, and calculate the new shifted centroids.
5. Remove any records that are closer to another shifted centroid and repeat until no records are lost.
6. Remove any records with negative silhouette values and repeat previous steps until no records are lost.

The resulting data sets represent the best-case scenario for individual cluster structures. Clusters exhibit complete separation and the additional step of requiring positive silhouette values means that all records are more similar to other records within their cluster than members of another cluster.

CREATING MULTIPLE CLUSTER STRUCTURE DATA

After generating data sets with single cluster structures, the next step in setting up the problem is to decide how they should be combined. The first question is with respect to independence, or overlap, of the two structures. The following panel shows the two extreme options.



The figure on the left represents highly correlated and overlapping segment structures, and that on the right maximizes independence. And one could vary the independence to come up with anything in between. The focus here will be on the independent extreme, which intuitively should give each structure the greatest opportunity to be clearly detected, and therefore the likelihood of the combined structure to emerge.

Given the strategy of independence, the next question is how to select structures to combine so that we can integrate characteristics that may influence which variables are selected from which structure. The full factorial single structure design has 32 cells, and there are 496 ways to choose 2 individual sets of characteristics from a set of 32.

It seems natural to think of this similar to a conjoint, with the data characteristics being the attributes and their specific values the levels. Then, an experimental design can be generated with the profiles defining the characteristics of each segment structure in the analysis data set.

Sawtooth Software's Lighthouse Studio was used to generate an experimental design that would guide the analysis data set construction step. The design included 10 versions, each with 16 tasks having 2 concepts. The balanced overlap option was selected to mitigate any potential dominating data characteristics. Processing time is a necessary consideration, so a between concept prohibition was imposed to prevent the combination of two 5 segment structures.

To be a little clearer, the table below shows how the experimental design is used to guide the analysis data set construction.

	Concept 1	Concept 2
Segments	3	5
Segment Size	Uneven	Even
Dimensions	5	3
Indicators per Dimension	5	5
Separation	Small	Large
	Data Set 1	Data Set 2
	⊕	
	Analysis Data	

Data sets defined by the concept profiles are randomized and concatenated to maximize independence. Each task in the design represents a single iteration in the analysis. The full design is simulated twice resulting in 320 (10 x 16 x 2) data sets analyzed, the results of which are presented next.

RESULTS

Identifying the Correct Number of Clusters

The first question of interest is whether or not the variable selection techniques help identify the correct number of clusters. Previous research suggests this is a difficult task under the best of circumstances (Chrzan and White 2021).

Recall the individual structures have either 3 or 5 segments and are combined to maximize independence. This means the true number of clusters for the combined structures is either 9 or

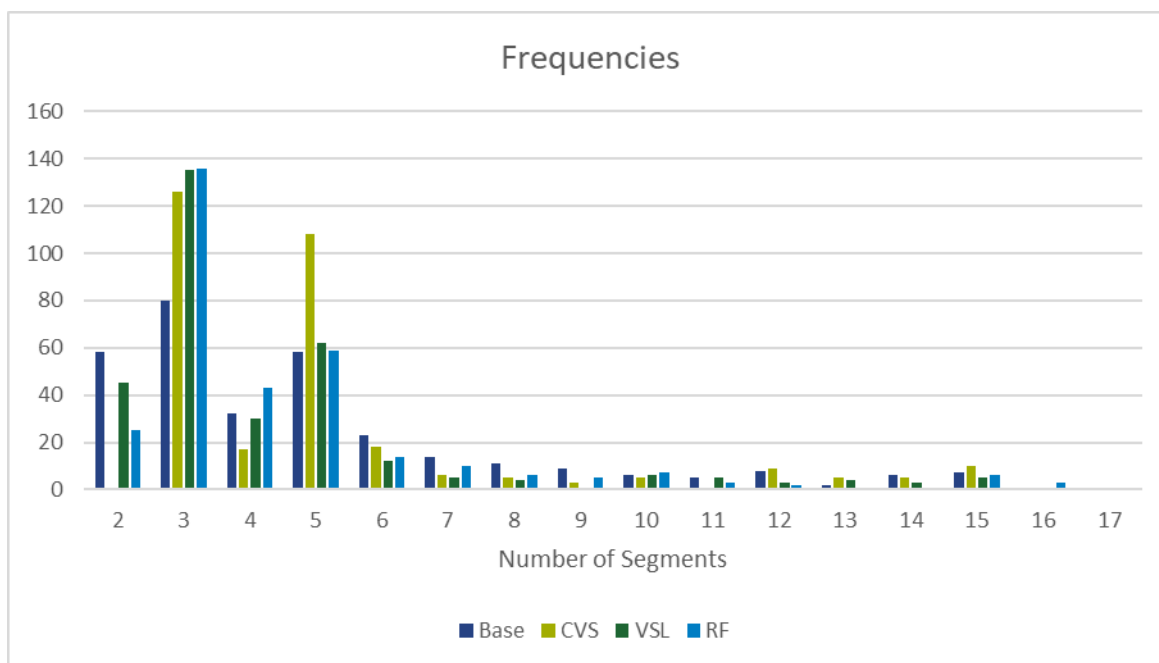
15. The 5x5 case is prohibited in the design, which also has the effect of suppressing the number of times a 3x3 analysis data set is considered, so most of the time the true joint distribution is comprised of 15 segments.

The R package mclust is used for the final segmentation using the identified basis variables and searching 2 to 17 cluster solutions to ensure coverage of the true number of clusters. The table below shows the performance of each technique at uncovering the complete joint number of clusters.

	Correct Number of Clusters Identified			
	Base	clustvarsel	VarSelLCM	randomForest
Count	7	10	5	6
Percent	2.2%	3.1%	1.6%	1.9%
Improvement	-	42.9%	-28.6%	-14.3%

Initial results are not promising for the joint segment structure. CVS marginally wins but the improvement over doing nothing is almost trivial. It appears highly unlikely that if there are multiple structures in your data that the statistics would suggest the correct solution, assuming of course that the joint structure is the ideal.

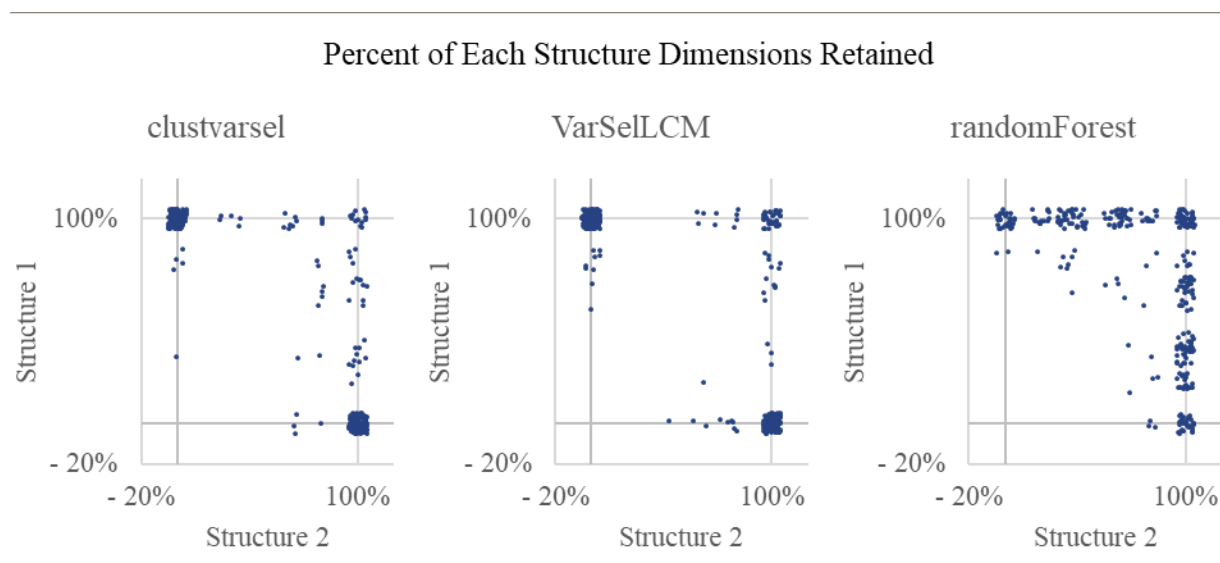
If the joint structure is this elusive, then the question turns to the marginals. Are variables being selected that reveal one structure clearly, or is partial information being retained from both, resulting in something less clear? The following chart looks at the number of clusters identified using the selected basis variables.



Individual cluster structures contained either 3 or 5 segments, so looking at the frequency of number identified starts to paint a picture that the selection techniques may be tending towards one or the other. This can be seen by the mass at 3 and 5, with CVS showing the greatest combined frequency on those two solutions, resulting in 3 or 5 clusters 72% of the time. The equivalent percentages for Base, VSL, and RF are 43%, 62%, and 61%, respectively.

Dimensional Retention

If variable selection is resulting in the identification of 3 or 5 clusters, the question is if they are pure solutions or a combination of the underlying structures. One way to start to answer this is to look at the dimensional retention of each technique to see which spaces emerge. The structure alignment with retained dimensions is presented in the panel below.



The vertical axis represents the percent of structure 1 dimensions retained and the horizontal that for structure 2. Each dot represents the results of a single iteration (task). Values range from 0% to 100% and are jittered to facilitate interpretation. A dimension is retained if at least one indicator for that dimension remains in the final set of selected variables.

The first thing to notice is that both CVS and VSL tend to zero in on the dimensionality of one structure and remove entirely that for the other. This is seen by the mass at the points (100%,0%) and (0%,100%) for CVS and VSL. The RF approach is much more likely to capture all the dimensionality for one structure and at least some from the other. This suggests that the 3 or 5 segment solutions obtained are indeed more often one of the unique structures.

In light of the results for finding the right number of clusters in conjunction with dimensional retention, the question of finding one structure versus another can be answered more definitively. A structure, in marginal terms, is defined as being correctly identified if the resulting number of clusters is the same as one of the single structures and at least half the dimensionality for that structure is retained. The table below is an example of success and failure based on this definition.

Task	Indicators Retained per Dimension										Clusters			
	Structure 1 Dims					Structure 2 Dims								Correct
	x1	x2	x3	x4	x5	y1	y2	y3	y4	y5	Str 1	Str 2	ID	
1	3	0	3	0	0	0	0	1	0	0	5	3	4	N
2	0	1	0	1	1	3	4	3	0	0	5	3	3	Y

These are two iterations, call them task 1 and 2, with dimensional and number of cluster results. Structure 1 is defined by dimensions x1 – x5, and structure 2 by y1 – y5. The retention portion of the table shows how many variables in each dimension were retained. In task 1 for

example, 3 indicators were retained from dimensions x1 and x3, and 1 from y3. Under the “Clusters” heading are the true number of clusters for each structure and the number that were identified in that iteration. Finally, the “Correct” column shows the outcome of the test.

Task 1 fails because only 4 clusters were identified and neither marginal structure consists of 4 segments. Task 2 is successful because it correctly identified 3 clusters and half of the dimensions retained were from the corresponding structure 2. The above examples are from the RF approach, which is more likely to retain cross-structure dimensionality. The definition of success is intended to allow for this dynamic because the way the data are generated does not prevent correlation between variables in competing structures.

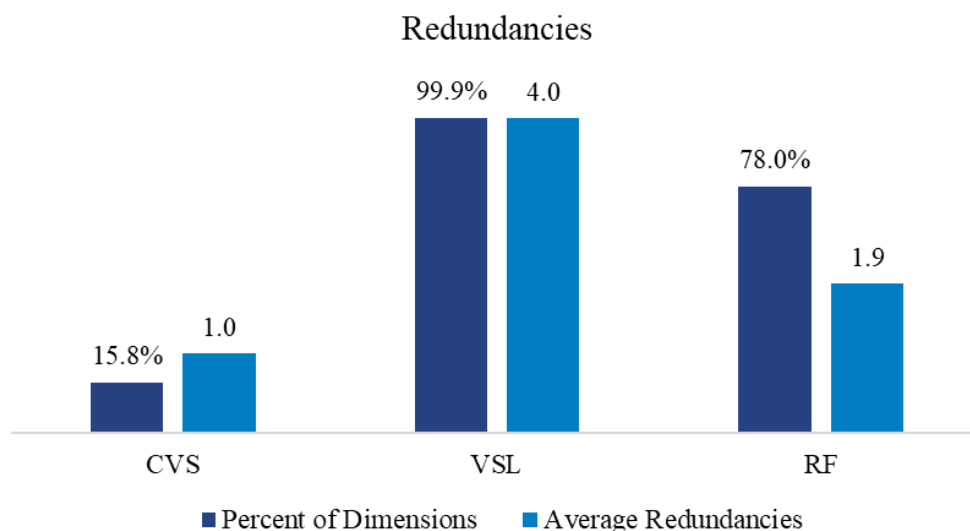
The definition of success is applied to all iterations where one structure contained 3 clusters and the other 5. This avoids the potential fuzziness associated with the handful of iterations comparing 3 cluster structures, although it would be easy enough to strengthen the definition to require a majority of dimensionality. This restriction removes just 8 iterations, and the results for the remaining 312 are presented below.

	Right Number		Correct Identification	
	Count	Percent	Count	Percent
clustvarsel	234	75.0%	222	71.2%
VarSelLCM	197	63.1%	172	55.1%
randomForest	195	62.5%	168	53.8%

The “Right Number” columns show how often either 3 or 5 clusters were identified using the reduced set of variables. The “Correct Identification” columns report the success of finding either 3 or 5 clusters and those being associated with the correct dimensional component as described above. CVS is clearly zeroing in on one cluster structure, and correctly identifying the right number of marginal segments. VSL and RF are on par with each other in this respect, clearly identifying the number of clusters for the right structure a little better than half the time.

Removing Redundancies

Related to the dimensionality question is that of removing redundancies. It appears that dimensionality is retained for one or the other structures, but are they retained cleanly or do correlated variables remain after the selection process? The chart below summarizes how well the tested techniques perform at removing redundancies.

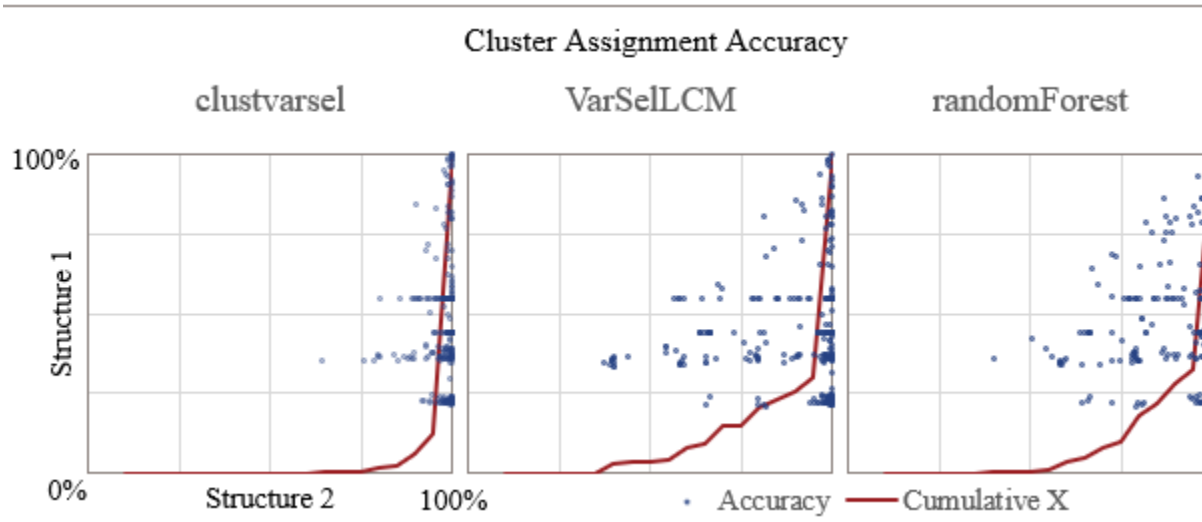


Percent of Dimensions refers to the average percent of the dimensions containing redundancies after the variable selection process, and Average Redundancies to the average number of redundant variables retained conditional on there being redundancies. On average, CVS retained redundancies in 15.8% of the dimensions identified, and when redundancies remained there was 1 on average. Because there are 5 indicators per dimension for iterations involving correlated variables, this means that for CVS there were 2 indicators on average for dimensions containing redundancies.

Both VSL and RF perform notably worse at removing redundancies. VSL essentially removes no correlated variables. This is obviously not an objective of the package, which is unfortunate because of the benefits of speed and the ability to handle multiple data types.

Accuracy

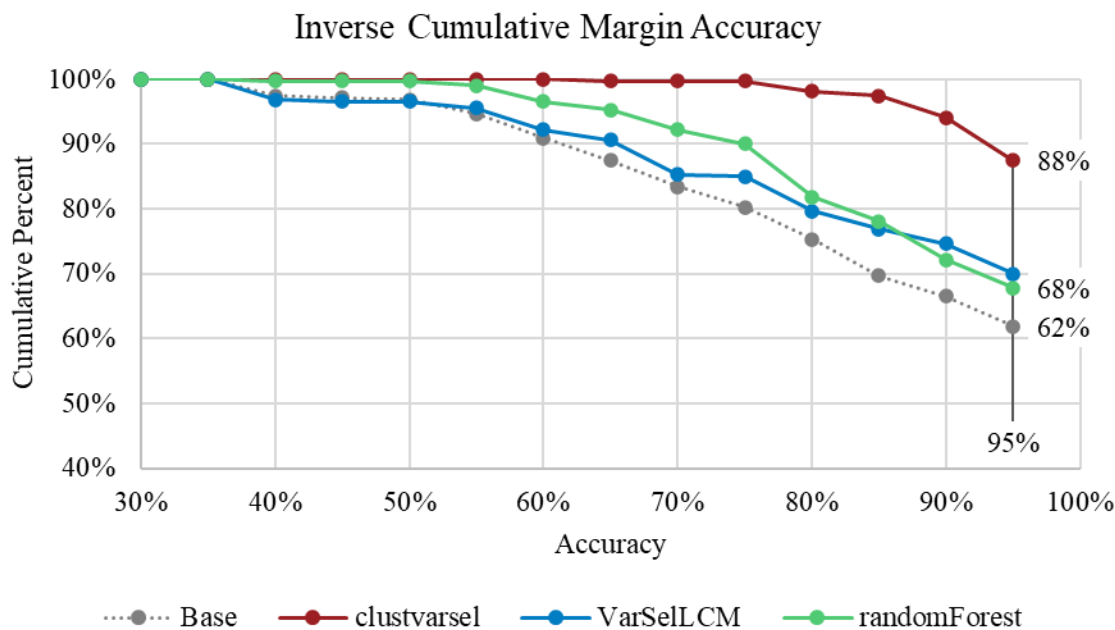
The results of the analysis thus far suggest variable selection techniques are picking up on one structure over the other more than a combination of the two. As a final analysis, accuracy of cluster assignments is considered in the context of marginal structures. How similar are the estimated cluster assignments to the known memberships for structure 1 and structure 2? The panel below plots accuracy for structure 1 versus structure 2.



Again, each dot represents a single iteration. The data were organized so that structure 2 is the one better recovered in terms of accuracy. The cumulative distribution of structure 2's accuracy is also charted as a way to better discern relative performance.

Given the question has turned from identifying the joint cluster structure to the margins, dots closer to the right and a lower cumulative distribution are desirable. VSL and RF appear to perform similarly at a glance, with CVS outperforming both. Recall that RF is more likely to retain dimensionality from the secondary structure so has conflicting information being included in the final set of basis variables, which in turn introduces noise into the segmentation algorithm, at least with respect to the primary structure. In addition, VSL did not remove redundancies, which has been seen to adversely affect segmentation algorithms.

It is easier to see the relative accuracy by looking at the interesting portions of the inverse cumulative distributions for the better recovered structure as in the following chart.



At any given level of accuracy this chart shows the percent of the time each technique performed at least that well. When interest is restricted to the emergent structure, CVS clearly outperforms either VSL or RF. This again is likely due to redundancies for VSL and cross-structure dimensionality for RF.

CONCLUSIONS

The automatic variable selection techniques considered in this paper appear to do a nice job of isolating a single cluster structure when two well-formed segment spaces coexist in our data. This is promising, as it suggests that when there are multiple structures, we have a decent chance of discovering at least one of the structures after effective variable selection. Of course, the risk is that the emergent structure, while statistically more apparent, is ultimately less interesting or useful from a business perspective than the structure that fails to emerge.

Similar to previous findings, CVS again rises to the top for variable selection. VSL, while showing a big improvement in terms of processing time, fails to remove any redundancies, which in the end is a likely contributor to the relatively poor performance of identifying the right number of clusters and accurately classifying observations.

The manual RF approach performs better than VSL at removing redundancies but is on par when it comes to accuracy. The interesting result of RF is the cross-structure dimensional retention. It seems to the author that there may be an opportunity to leverage this in conjunction with CVS and another analysis to be determined as a way to uncover the existence of multiple segment structures.

SUGGESTIONS FOR FUTURE RESEARCH

This paper has considered the very special case of independent complete cluster structures on our ability to effectively select basis variables in segmentation. The assumption of independence in the sense of maximizing the cross-segment dimensionality is an edge scenario. What about the more likely case where the structures are more correlated? Do the basis variables for one structure serve as masking variables for the other when independence is maximized? As we move to the other extreme of correlated cluster spaces, do the basis variables for one become redundancies for the other?



Joseph White

REFERENCES

Andrews, J. L., and P.D. McNicholas (2013) “Variable selection for clustering and classification,” *Journal of Classification*, **31**(2), 136–153.

- Brusco, M. (2004) “Clustering binary data in the presence of masking variables,” *Psychological Methods*, **9**(4): 510–523.
- Chrzan, K. and J. White (2022) “Variable selection in segmentation,” *Sawtooth Software Conference Proceedings*, 207–218.
- Chrzan, K. and J. White (2021) “Replication of known segment structure and membership,” *Sawtooth Software Conference Proceedings*, 217–226.
- Dolnicar, S., B. Grün, and F. Leisch (2018) *Market Segmentation Analysis: Understanding It, Doing It, and Making It Useful*. Singapore: Springer.
- Dolnicar, S., B. Grün and F. Leisch (2016) “Increasing sample size compensates for data problems in segmentation studies,” *Journal of Business Research*, **69**: 992–999.
- Formann, A. (1984) *Die Latent-Class-Analyse: Einführung in die Theorie und Anwendung*. Beltz, Weinheim.
- Hartigan, J.A. (1972) “Direct clustering of a data matrix,” *Journal of the American Statistical Association*, **67**(337): 123–129.
- Kaiser, S. (2011) *Biclustering: methods, software and application*. Ph.D. thesis, Department of Statistics, Ludwig-Maximilians-Universität München, Munich.
<https://edoc.ub.uni-muenchen.de/13073/>
- Milligan, G.W. (1980). “An examination of six types of the effect of six types of error perturbation on fifteen clustering algorithms.” *Psychometrika*, **45**, 325–342.
- Milligan, G.W. and M.C. Cooper (1985) “An examination of procedures for determining the number of clusters in a data set,” *Psychometrika*, **50**: 159–179.
- Nowakowska, E. and J. Retzer (2021) “BiCluster identification and profiling,” *Sawtooth Software Conference Proceedings*, 227–238.
- Qiu, W. and H. Joe (2020) “clusterGeneration: random cluster generation (with specified degree of separation).”
<https://cran.r-project.org/web/packages/clusterGeneration/clusterGeneration.pdf>.
- Sawtooth Software, Inc. (2013) “CCEA V3,” downloaded from
<https://sawtoothsoftware.com/resources/software-downloads/convergent-cluster-ensemble-analysis>. Accessed 10/18/2021.
- Scrucca, L. and A.E. Raferty (2018) “clustvarsel: A package implementing variable selection for Gaussian model-based clustering in R,” *Journal of Statistical Software*, **84**(1): 1–28.
- Yiu, T. “The Curse of Dimensionality: Why high dimensional data can be so troublesome,” *Towards Data Science*, July 20, 2019, <https://towardsdatascience.com/the-curse-of-dimensionality-50dc6e49aa1e>. Accessed 10/18/2021.

DESIGN AND ESTIMATION IN A CBC STUDY WITH ADDITIVE BINARY ATTRIBUTES AND PRICE

TOMMASO GENNARI
ANALYTICS WITH PURPOSE

ABSTRACT

We designed and modelled a CBC study with 12 binary attributes and price. Designs and models of this type can be used when the product subject of the study is composed of several features, that can be present or absent. This is a case that can be relatively common for marketers to address. Standard CBC designs and models are usually not fit to this case, because of several potential issues, among which is the possible correlation between predictors of the choice, implying diminishing return effects. Even if this is a complex and potentially common type of study, it seems that we lack a shared way of addressing it in our community. Solutions involving sophisticated modelling have been proposed, but they might be out of reach for the normal researcher, precisely due to their sophistication. We illustrate here a method that worked in our case. This method avoids sophisticated modelling but requires ad hoc designing and the use of a few advanced CBC features. We hope that as a community of researchers we can solidify methods to address such types of studies.

BUSINESS PROBLEM AND OBJECTIVE OF THE ANALYSIS

The case presented in this paper refers to a project executed in 2022. Any reference to the specific subject of this project, the client and the industry are confidential. However, the case presented here represents a common research question from insight managers.

The client asked to solve a complex price/feature problem. Their potential product was made up of several features, that could be present or absent, and their objectives were: 1) set up bundled tiers of product configurations made up of features, 2) starting from a simple configuration to a more complex, 3) where each successive bundled tier includes all features present in the simpler tiers.

Figure 1: Ideal Bundled Tiered Configuration Objective of this Project

	Tier 1	Tier 2	Tier 3	Tier 4
Feature 1	x	x	x	x
Feature 2	x	x	x	x
Feature 3	x	x	x	x
Feature 4		x	x	x
Feature 5		x	x	x
Feature 6		x	x	x
Feature 7			x	x
Feature 8			x	x
Feature 9			x	x
Feature 10				x
Feature 11				x
Feature 12				x
Cost in USD:	0	70	150	200

There was not a specific optimization objective, but rather a generic request for identifying the ideal number of tiers and price of each tier. The client had initial assumptions regarding price: they had an indication of how much the higher tier (the one including all features) would need to cost, and that the lowest first tier (the one with fewer features) should be free.

DESIGN: OVERVIEW

With this brief, we set out to identify what we want to learn from a sample of potential customers with a survey. Ideally, we would like to know: 1) their preference for each feature, 2) their preference for any given subset of features vs. other subset of features (tiers), 3) which tier would they choose when faced with the choice of a specific tier, 4) if they would choose a tier at all.

We considered and assumed that, price being constant, customers would prefer to buy a tier with **more features** compared to a tier with fewer features. We called **n** the number of features in a tier/alternative; n is a key element of the research.

We considered and assumed that tiers with more features should be more expensive.

We considered different types of conjoint exercises (e.g., adaptive, partial profiles) before deciding to focus on a standard CBC set up, where we collected important information with an ad hoc design, so that we had the right information to feed into a model later.

DESIGN: FEATURES AND SCREENS

After the “usual” negotiation with the client we ended up with **12 features** to use. This is not a huge number, but still a number which requires quite a lot of screen space to be visualized.

The structure of a screen for a respondent is shown in Figure 2. Using the standard CBC set up in Lighthouse allowed the screen to be mobile friendly.

Figure 2: Example of Screen Set Up

	Product A	Product B	Product C	Product D
Feature 1				x
Feature 2	x			x
Feature 3	x		x	x
Feature 4	x	x		
Feature 5		x	x	x
Feature 6	x		x	x
Feature 7		x		x
Feature 8		x		x
Feature 9		x		x
Feature 10		x		
Feature 11		x		
Feature 12		x		
Price	\$0	\$200	\$25	\$180

We decided to keep the same order of the features across the screens, to facilitate the reading of each screen for respondents. The order of the features was logically derived from their content. A hover box was present to allow the full description of each feature.

The prices were not ordered by column, like shown in the example in Figure 2.

Alternatives that were fully included in other alternatives on the same screen, like Product C being fully included in Products D in Figure 2, were designed to have a price lower than the alternative that includes them.

We designed 10 screens only for each respondent, trying to avoid too many of them, fully aware that each screen requires a lot of effort.

A Dual None option was present, so that at each screen, respondents were asked if they would really buy the option they chose as their favorite.

DESIGN: PRICE

Because of our assumptions of price being higher for tiers with more features, we designed a price structure conditional to the number of features, as illustrated in Figure 3.

Figure 3: Conditional Price Structure

# features	Prices (USD):				
	low	mid low	mid	mid high	high
1	0	6	17	25	33
2	0	13	33	49	65
3	0	22	50	72	94
4	0	32	67	94	121
5	5	44	83	114	145
6	15	58	100	134	168
7	29	73	117	153	189
8	45	89	133	170	207
9	64	107	150	187	224
10	86	127	167	202	238
11	112	148	183	217	250
12	140	170	200	230	260

This structure includes enough free options when the number of features is low, and a good overlapping of price levels between different numbers of features: we wanted, for example, the lower level of prices for alternatives with 7 features to be lower than the higher level of prices of the alternatives with 6 features, and so on. This structure was built on the client's expectation that the tier with all features should be priced at around \$200.

DESIGN: AVOIDING THE NORMALITY TRAP

We have designed alternatives made up of 12 possible features and a price. For this reason, if we were to adopt a pure uniform random design, we would have very few alternatives with a small number or a big number of features, and a lot with 5 to 7 features. This would be a problem if we wanted to use the choice data collected to estimate the value or customers of the number of features, or to use in the market simulator tiers made of a small or a big number of features.

For this reason, we have applied multipliers conditional to the number of features n to the design generation, as illustrated at Figure 4. This allowed us to collect enough information about choices, also when n is small or big.

Figure 4: Distribution of the designed alternatives by number of features with uniform random generation, multiplier we used in the random generation, and actual distribution of alternatives by number of features we have used in our design.

number of features	average times shown to a respondent if random design used	multiplier	average times shown to a respondent in our design
1	0.1	25	2.9
2	0.6	5	3.2
3	2.1	1.5	3.2
4	4.8	0.5	2.4
5	7.7	0.7	5.4
6	9.0	0.7	6.3
7	7.7	0.5	3.9
8	4.8	0.5	2.4
9	2.1	1	2.1
10	0.6	5	3.2
11	0.1	25	2.9
12	0.0	200	2.0

This precaution is also described by Lattery (2013) in a paper he wrote following the Sawtooth Software Conference of October 2013, where he describes a case similar to the ones we are writing about here.

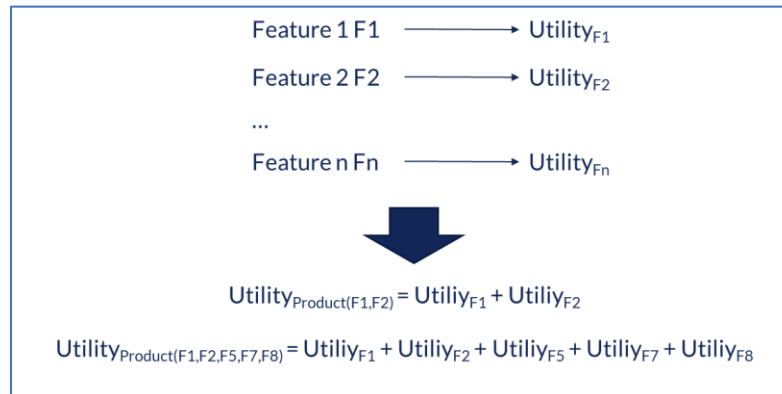
THE ISSUE OF DIMINISHING RETURNS

Before passing on to how we modelled data collected with such a design, we need to explain why and how diminishing returns can be an issue in this context.

Intuitively, the issue lays in the fact that, normally, preferences for levels of attributes (in this case, features) are estimated with coefficients of a regression model, and these coefficients are added up cumulatively to estimate the utility of a product made up of those levels/features. In this sense, the utility/importance of each feature is independent from each other and additive in the estimation of the total utility of an alternative/product.

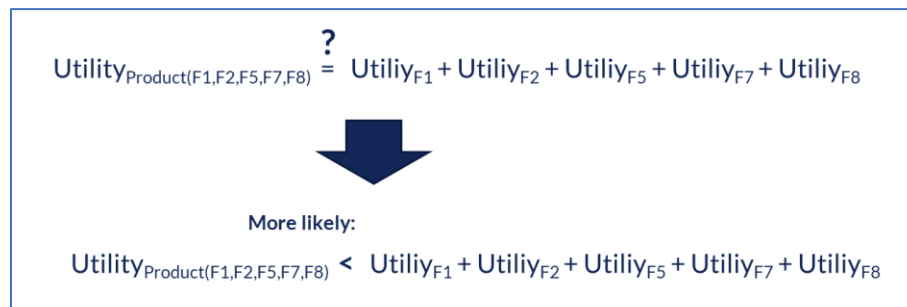
Figure 5: “Standard” typical additive representation of how purchasers/consumers are assumed to have a certain level of preference/utility for each feature of the product/alternative, and how these combine together to build up the total utility of the product/alternative.

(This representation is illustrative only and it does not pretend to be mathematically complete or precise.)



Now, when we have a relatively long list of features, like in our case, one might argue that this additive property does not hold. This is because the value of the utility of a single feature might be different when we use it to define a product with fewer other features, compared to when we use the same feature to define a product/alternative with many more features.

Figure 6: Illustration of the presence of diminishing returns— it is expected that the utility of a product made of many features is smaller than the sum of the individual utilities of the single features.



We would expect the individual contribution of a feature to be bigger when it is part of a product/alternative with fewer features, and smaller when the same feature defines a product/alternative together with many other features. In this sense, we would expect a level of utility of a feature being smaller when n (the number of features in an alternative/product) gets bigger. This is why we can talk about diminishing returns.

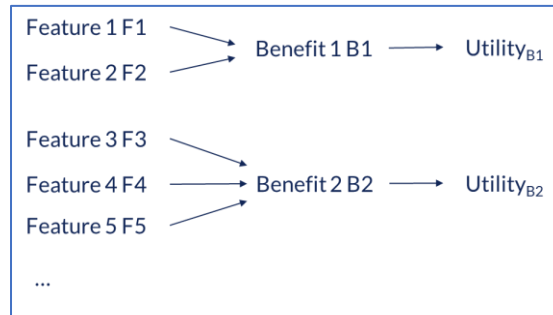
A very interesting discussion of this problem is presented in Lattery (2013). In this paper, Lattery describes how other researchers address this issue, and proposes the use of nested logit modelling to incorporate diminishing returns in the model.

Lattery shows with actual survey data that, if this issue is not addressed properly, there will be an overestimation of the shares of preferences for alternatives/products made of more features.

Kim et al. (2017) develops further the use of nested logit modelling in similar cases. They explain that when products are defined by many binary features (like in our case), the carriers of independent utilities for the purchaser/consumer might actually be some unobserved benefits, to which each feature can belong.

They also explain that the reason why we see diminishing return is not that purchasers/consumers “just” consider less a feature only because it is added as $n+1$ on top of products with already n other features, but because purchasers/consumers derive utility not from the features per se, but from the perceived benefits that the features carry. In this sense, even if a product might be composed of n features, the benefits the purchaser/consumer perceive might be only k , with $k < n$ or even $k \ll n$. In our case, even if we have 12 features, the benefits might have been 3 or 4. In this sense, when a product is already covering a specific benefit with a feature carrying it, the addition of a second feature carrying the same benefit will improve only by a little the total utility of the alternative/product for this consumer/purchaser.

Figure 7: Illustration of the logical structure proposed by Kim et al. (2017) to explain why diminishing returns are observed.



After describing why we can see diminishing returns in such cases, Kim et al. (2017) explains how nested modelling can be used to address that.

When we originally worked on the project presented in this paper, we were unaware of the two papers discussed above. We believe that nested logit modelling is not easily usable by the average conjoint researcher. In this sense, we still consider the design and the modelling that we present to be a viable way of addressing the challenges presented by conjoint with binary attributes.

MODEL ESTIMATION

We produced predictive models of the choices in a relatively standard way, using the software CBC/HB by Sawtooth Software and considered price as a linear predictor in USD. The only “deviation” from a standard model was that we tested the use of an additional predictor, i.e., the number of features n in an alternative. As you can see from Figure 8, these models had a relatively low in-sample predictive power as measured by the RLH measure.

Figure 8: First Modelling Iteration Using a Relatively Standard Approach

Iteration:	Predictors used in the successive iterations of modelling in CBC HB:						RLH
	Features	Number of features	Beta of number of features constrained to be positive	Linear price	Beta of linear price constrained to be negative	Number of parameters	
0	x	x		x		13	0.44
1	x			x		12	0.43
2	x			x	x	12	0.39
3	x	x	x	x	x	13	0.39
4		x	x	x	x	2	0.31
5	x					11	0.38
Number of parameters	11	1		1			

However, more than the low RLH, the most worrisome characteristic of these models was the de facto absence of price sensitivity. We realized that this characteristic was a direct consequence of these models not being aligned to the design we used. In our design, price is naturally correlated to the number of features. Our model needs to include this characteristic of the design in its build.

To address this issue, we introduced a non-standard predictor, i.e., the full interaction of level of price and number of features. As illustrated in Figure 9, this predictor is a categorical variable with 60 levels, because we used 5 levels of price and 12 number of features ($5 \times 12 = 60$)

Figure 9: Levels of the categorical predictor “Price by Number of Features,” and their assumed constraints.

# features	Prices:									
	1 low		2 mid low		3 mid		4 mid high		5 high	
1	B(n=1,p=1)	>	B(n=1,p=2)	>	B(n=1,p=3)	>	B(n=1,p=4)	>	B(n=1,p=5)	
2	B(n=2,p=1)	>	B(n=2,p=2)	>	B(n=2,p=3)	>	B(n=2,p=4)	>	B(n=2,p=5)	
3	B(n=3,p=1)	>	B(n=3,p=2)	>	B(n=3,p=3)	>	B(n=3,p=4)	>	B(n=3,p=5)	
4	B(n=4,p=1)	>	B(n=4,p=2)	>	B(n=4,p=3)	>	B(n=4,p=4)	>	B(n=4,p=5)	
5	B(n=5,p=1)	>	B(n=5,p=2)	>	B(n=5,p=3)	>	B(n=5,p=4)	>	B(n=5,p=5)	
6	B(n=6,p=1)	>	B(n=6,p=2)	>	B(n=6,p=3)	>	B(n=6,p=4)	>	B(n=6,p=5)	
7	B(n=7,p=1)	>	B(n=7,p=2)	>	B(n=7,p=3)	>	B(n=7,p=4)	>	B(n=7,p=5)	
8	B(n=8,p=1)	>	B(n=8,p=2)	>	B(n=8,p=3)	>	B(n=8,p=4)	>	B(n=8,p=5)	
9	B(n=9,p=1)	>	B(n=9,p=2)	>	B(n=9,p=3)	>	B(n=9,p=4)	>	B(n=9,p=5)	
10	B(n=10,p=1)	>	B(n=10,p=2)	>	B(n=10,p=3)	>	B(n=10,p=4)	>	B(n=10,p=5)	
11	B(n=11,p=1)	>	B(n=11,p=2)	>	B(n=11,p=3)	>	B(n=11,p=4)	>	B(n=11,p=5)	
12	B(n=12,p=1)	>	B(n=12,p=2)	>	B(n=12,p=3)	>	B(n=12,p=4)	>	B(n=12,p=5)	

After introducing this predictor, the fit of the models changed dramatically, as you can see in Figure 10. RHL was higher, and the utilities from the model were expressive of price sensitivity, as we illustrate in the next section.

Figure 10: Full set of the models tested.

Models 0 to 5 only include the number of features n as a non-standard predictor. Models 6 to 8 include the full interaction between price and number features among the predictors.

Iteration:	Predictors used in the successive iterations of modelling in CBC HB:								RLH
	Features	Number of features	Beta of number of features constrained to be positive	Linear price	Beta of linear price constrained to be negative	Price levels * number of features	Betas of price levels decreasing within number of features	Covariates	
0	x	x		x					0.44
1	x			x					0.43
2	x			x	x				0.39
3	x	x	x	x	x				0.39
4		x	x	x	x				0.31
5	x								0.38
6	x					x			0.68
7	x					x	x		0.53
8	x					x	x	x	0.56
Number of parameters	11	1		1		59			

Model 6 included among the predictors:

- The 12 binary variables indicating the presence or the absence of the features in the alternative.
- The full interaction of price and number of features.

Model 7 included among the predictors:

- The 12 binary variables indicating the presence or the absence of the features in the alternative.
- The full interaction of price and number of features.
- The level indicating different prices were constrained, as illustrated at Figure 9, following the assumptions that purchasers/consumers would on average prefer a cheaper alternative, while the number of features is the same. We assumed that a more expensive alternative, n being constant, should be chosen only if the set of utilities of its features counterbalances this price sensitivity; in other words, consumers would choose a most expensive alternative with the same number of features, only if they have a distinct stronger preference for the features included in this alternative.

Model 8 included the same predictors and constraints of Model 7, plus demographic information.

We chose Model 7 from Figure 10 as the model to use in the market simulator, for 3 reasons:

- RHL is not extremely high, as in Model 6, and we considered that this might have been caused by a regularization effect of the constraints. In other words, we suspected Model 6 to be more likely affected by overfitting compared to Model 7.
- The analysis of the utilities as shown in the next section is in line with our expectations.
- The market simulator made with Model 7 has behavior matching expectations and is able to help our client to achieve their objectives.

ANALYSIS OF THE UTILITIES

The most convincing piece of evidence that made us choose Model 7 is reported in Figure 11. This is the average utility by number of features and price. We considered this pattern to be in line with our expectations. These are the utilities from Model 7 described in the previous section. This model used as predictors an indicator for each feature and the full interaction between price and number of features, the price utilities being constrained within each number of features.

Figure 11: Average utility of alternatives by number of features it is composed of and price (Model 7).

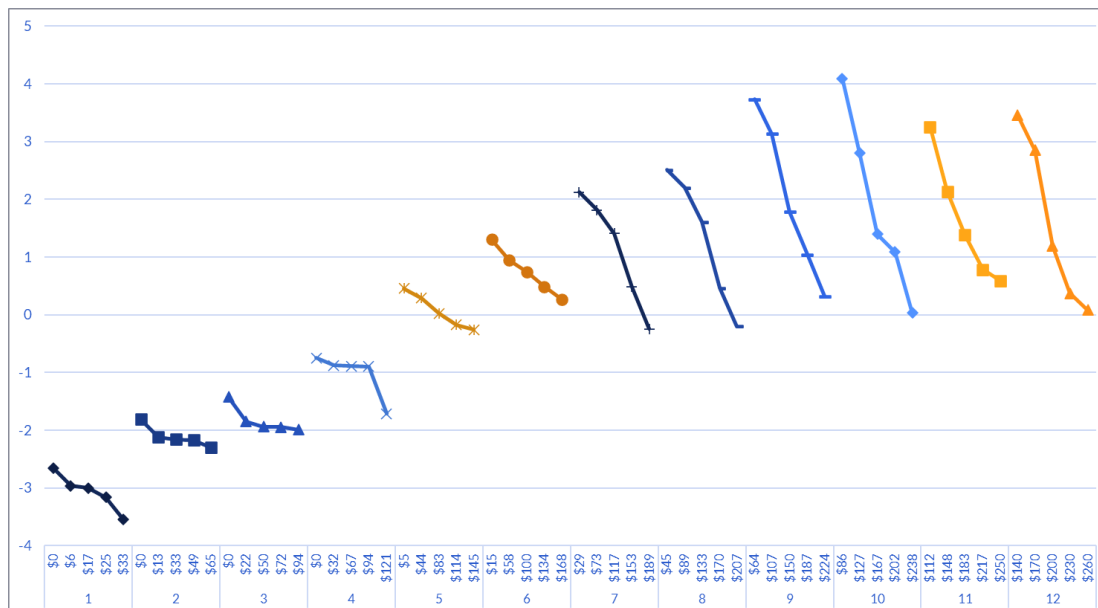


Figure 11 shows price sensitivity as we expect it within each number of features from $n=5$ up. Also it shows that in general when n is increased by one, respondents are more likely to accept an increase of price.

Figure 11 shows also that price sensitivity is not strong for $n < 5$. This might be due to a series of reasons, of which our prime suspect is that the variation of prices we used in the design for $n < 5$ is not large in absolute terms, because we approximately applied percentage price variations. We might argue that if we were to use larger price variations for n small, we could have collected better information to assess price sensitivity.

Figure 11 also shows decreasing marginal returns when n grows. The average increase of the utilities is decreasing when n grows, which is something we would expect after the discussion in the previous section dedicated to the review of Lattery (2013) and Kim et al. (2017).

MARKET SIMULATOR AND OUTCOME FOR THE CLIENT

As mentioned in the previous sections, we chose Model 7 as the model with which to build the market simulator. The results were in line with our and our client's expectations, and the market simulator was instrumental in helping our client decide tiered bundles and their prices.

To give more details about the outcome and the implementation of the market simulator:

- To decide which features to bundle in each tier, we split the sample by price sensitivity, and analyzed the level of importance of the features by this split. We then decided to build a first “entry level” free tier with the features more important to the more price sensitive respondent, who would not pay for more anyway, but at least have a way to be “hooked” into the product. Eventually, we assigned to the last and most expensive tier the features more important for the less price sensitive respondents.
- Because we wanted to use price as a linear input in the market simulator, we transformed the utilities measured on the 5 price points for each n via linear interpolation.

DISCUSSION

We recognize that the predictive model could have been estimated differently, but we are confident that the version we used helped to give our client the right insights into the expected market behavior. This is due to the factors discussed throughout the paper: high in-sample predictivity measured by RLH, alignment with the expectations, meaningful patterns of utilities by n and price, meaningful price sensitivity, presence of diminishing marginal returns.

Some of the suggestions given, about what could have been done differently, during the Analytics and Insights Summit:

- More exploratory analysis of the model for $n < 5$.
- A more parsimonious model (a model with fewer predictors) once the pattern at Figure 11 was assessed.
- Testing models with different functional dependencies between predictors and choices.
- And, of course, using the model described by Kim et al. (2017).

One of the critical aspects that makes our modelling diverge from that proposed in Lattery (2013) and Kim et al. (2017), is that we have modelled diminishing returns that are independent from the features/attributes, so that the reduction of the increase in utility only depends on n (number of features). Lattery (2013) and Kim et al. (2017) propose a model in which the decrease in marginal utility depends on which feature is added as $n+1$ feature.

TAKEAWAYS

Because of the satisfactory outcome of our analysis, we believe that we have offered a blueprint for designing and modelling complex conjoint cases like the ones illustrated here. This is a summary of what we recommend doing in these cases:

- Create an ad hoc design making sure that you collect the information necessary for a predictive model to be estimated, among which are:
 - Enough alternatives with a small or large number of features.
 - Enough price variation especially at the lowest level of price.
- At the modelling stage:
 - Test a model with full interaction of price and number of features as a predictor.
 - Explore price sensitivity.
 - If possible keep a small number of predictors, reducing the complexity of full interaction of price and number of features.

CONCLUSIONS

The use of many binary attributes, i.e., features, in conjoint analysis introduces complications in the design, and possibly overestimation of the share of preferences for products made of many of these features (diminishing returns). The use of nested logit models can help in overcoming these issues, but this type of modelling is not within the reach of the average researcher. We proposed here an alternative approach that simplifies some assumptions behind the modelling of diminishing returns (specifically, considering the diminishing return only function of the number of features, and not of the particular features themselves), but it is implementable using standard software used for conjoint analysis.



Tommaso Gennari

REFERENCES

- Lattery, K. (2013). When $U = BX$ is not Enough: Modeling Diminishing Returns Among Correlated Conjoint Attributes, *Proceeding of the Sawtooth Software Conference, October 2013*.
- Kim, D.S, Bailey, R.A, Hardt, N. and Allenby, G.M. (2017). Benefit-Based Conjoint Analysis, *Marketing Science*, 36:1, 54–69

BUILDING DESIGNS FOR INDIVIDUAL-LEVEL ESTIMATION: CONSIDERATIONS, IMPLICATIONS AND NEW TOOLS

MEGAN PEITZ
TREVOR OLSEN
NUMERIOUS INC.

ABSTRACT

Choice-based conjoint (CBC) experiments are widely used to understand consumer preferences and willingness to pay for different product features. One important consideration in designing CBC experiments is the balance of attribute levels across the design. Implementing this strategy seeks to give every level an equal chance to influence the respondent's decision in the conjoint design and can work in the majority of cases. However, the authors of this paper were interested in revisiting the work of Huber and Zwerina (1996) to determine if utility balanced designs, a design strategy that trades off on level balance while optimizing which alternatives are paired against each other within tasks, could result in better predictions at the individual level. This paper sets out to explore several different methods of optimizing designs and offers access to an open-source package, built in Julia by the Numerious team, to leverage these different design strategies in the future.

The results from this paper show that utility balanced designs perform well in predicting data from both utility balanced and non-utility balanced designs, and that respondents do not seem to be fatigued by utility balanced designs. This would suggest that utility balanced designs could be a successful strategy depending on the attributes and levels being tested. However, we must caution the user of utility balanced designs as some design strategies may result in sparse data at the interaction level. We also believe that further research is needed to understand the differences in willingness to pay estimates between utility balanced designs and traditional, level balanced designs. It should also be noted that there are several different strategies for creating efficient designs as well as other packages used to generate the designs outside of what is mentioned in this paper. See References for more details.

BACKGROUND

According to Rossi, Allenby, McCulloch (2005) the greatest challenges in marketing are to understand the heterogeneity in preferences. This is why marketing practitioners prefer unit-level hierarchical Bayesian estimates.

To uncover those unit-level estimates, we are often taught to build designs that are level balanced (i.e., within each attribute, each level appears an equal number of times). Implementing this strategy seeks to give every level an equal chance to influence the respondent's decision in the conjoint design and can work in the majority of cases.

But McFadden (1974) shows that the estimated utilities from the model depend not just on which concepts are included in the design, but which concepts are paired against each other. The multinomial logit model (MNL) assumes part-worth utilities are independent of each other (i.e., preference for one level does not depend on the preference for another level). However, certain combinations of attributes and levels can affect the distribution of preferences among respondents. For example, it would not be surprising to see a Ferrari at a \$250K price point and a

Chevy Volt at \$25K. But it wouldn't make much sense to see a Chevy Volt at \$125K—so why would we waste observations on combinations that aren't relevant? Because of circumstances like this, we think good designs should not just be a matter of level balance across alternatives, but designs should also be dependent on which alternatives are paired against each other within the CBC tasks.

One could also argue that designs that optimize for the principles above could result in smoothing over the unit-level estimates (Bayesian Shrinkage), muting the individual level preferences and potentially resulting in poorer insights into the true heterogeneity of the marketplace.

So, what is a researcher to do? And is it really that big of a deal if we continue building designs according to these principles?

NON-LEVEL BALANCED AND UTILITY BALANCED DESIGNS

One solution is to use a design that is not level balanced. A non-level balanced design could result in some levels appearing significantly more frequently than others, and some pairs of levels appearing more or less frequently than others. However, this type of design may be useful when the relationships between the attributes and levels are complex, and/or where it's important to test specific interactions between the attributes.

Utility balanced designs are one alternative for a non-level balanced design and it has been shown (Huber and Zwerina, 1996) that designs which include utility balance as one criterion can improve the understanding of aggregate effects.

Utility balanced designs are conjoint designs in which the total utility of each concept shown within a task is as even as possible. Figure 1.1 shows the difference between what a level balanced design might look like versus a utility balanced design.

Figure 1.1: Level Balanced vs. Utility Balanced Designs



In reviewing Figure 1.1, we can see the imbalance of levels within the utility balanced design. This imbalance can cause apprehension among many researchers who would warn that if

the prior understanding is misspecified (i.e., people will spend less for a Ferrari and more for a Chevrolet vs. more for a Ferrari and less for a Chevrolet), then the resulting alternative comparisons will be less efficient than a design built with the prior at 0. However, if using the R idfix package (Traets, F., Sanchez, D.G. and Vandebroek, M. 2020), one can balance the prior utilities from a Bayesian perspective. In this approach, users can specify a full distribution of prior knowledge to balance the utility within tasks. This approach should avoid misspecification since more uncertainty is being incorporated into the design. Therefore, we will field a study with a level balanced design for n=50 completes and capture individual preferences with a hierarchical Bayesian model to seed our utility balanced designs.

Another concern around utility balanced designs is that the CBC tasks become too difficult and respondents become fatigued if the choices are too hard. And if the choices are too hard, respondent error outweighs the added design efficiency. To address this concern, we will ask respondents to rate the designs on measures like easy vs. hard, long vs. short, as well as explore completion time, drop-off rates, percentage of bad actors (i.e., cheaters) and ultimately the error around their responses by examining both within-sample and out-of-sample holdouts.

METHODOLOGY

We conducted an online survey about TVs with over 3,500 real respondents. Attributes and levels of the CBC exercise are shown in Figure 2.1 and a screenshot of the conjoint exercise is shown in Figure 2.2.

Figure 2.1: Conjoint Attributes and Levels

Brand	Resolution	Screen Size	Refresh Rate	Screen Technology	HDMI Ports	Price
Sony	4K	55 inches	60 Hz	LED LCD	3	\$450
LG	8K	65 inches	120 Hz	QLED	4	\$800
Vizio		75 inches		OLED		\$1,300
Samsung						\$1,900
TCL						\$2,700

Figure 2.2: Conjoint Exercise Screenshot

If these were your only options for HD TVs, which would you choose?
(1 of 14)

	Sony	Vizio	TCL	TCL
Brand:	Sony	Vizio	TCL	TCL
Resolution:	4K	4K	8K	8K
Screen Size:	55 inches	55 inches	65 inches	75 inches
Refresh Rate:	60 Hz	60 Hz	120 Hz	120 Hz
Screen Technology:	QLED	LED LCD	OLED	LED LCD
HDMI Ports:	3	3	4	4
Price:	\$1,300	\$450	\$800	\$1,900
	<input checked="" type="button" value="✓"/>	Select	Select	Select

Given what you know about the market, would you really buy the HD TV you chose above?

The conjoint experiment consisted of 14 choice tasks, each containing 4 product profiles and a dual-response none alternative. 12 choice tasks were designed by the algorithm (more details below) and 2 of the choice tasks were fixed, meaning that all respondents saw the same combinations of attributes and levels on two screens.

Respondents were assigned to one of eight of the following cells (Figure 2.3) with approximately n=450 completes per cell. All cells had 300 versions of the design except Cell 6 which is a 1 version design used for out-of-sample holdout validation.

Figure 2.3: Respondent Cells

	Cell 1	Lighthouse Balanced and Overlap	Level Balanced
	Cell 2	Discover 2.0	Level Balanced
	Cell 3	Julia no balance penalty	Level Balanced
	Cell 4	Julia balance penalty	Level Balanced
	Cell 5A	Julia utility balanced with a balanced penalty (based on N=50 from cell 4)	Not Level Balanced
	Cell 5B	Julia utility balanced with a balanced penalty (based on N=50 from cell 4 and utilities are 50% of their original size)	Not Level Balanced
	Cell 5C	Julia utility balanced with NO balanced penalty (based on N=50 from cell 4)	Not Level Balanced
	Cell 6	Lighthouse Balanced and Overlap (1 version - used for out-of-sample holdouts only)	Level Balanced

Sawtooth Software is leveraged to create the designs in Cells 1, 2, and 6. You can read more about the design algorithms used in Lighthouse and Discover on Sawtooth Software’s website (www.sawtoothsoftware.com). The Numerous cells were built in Julia (<https://julialang.org/>), an open-source programming language particularly suited for computational math.

The five Julia cells vary based on whether there is a balance penalty, whether there is a utility prior and whether that utility prior will be scaled.

The goal of the cells that have no balance penalty is to minimize D-error versus those with a balance penalty will trade-off minimizing the D-error in order to obtain more level balance. The three cells that have a utility prior leverage a hierarchical Bayesian model built from the first n=50 to respond to Cell 4. Then, within those utility prior cells, we will either trust the priors entirely and allow them to be 100% of their original size or we will shrink them to 50% of their original size. A high-level overview is below in Figure 2.4.

Figure 2.4: Overview of Julia/Numerous Design Cells

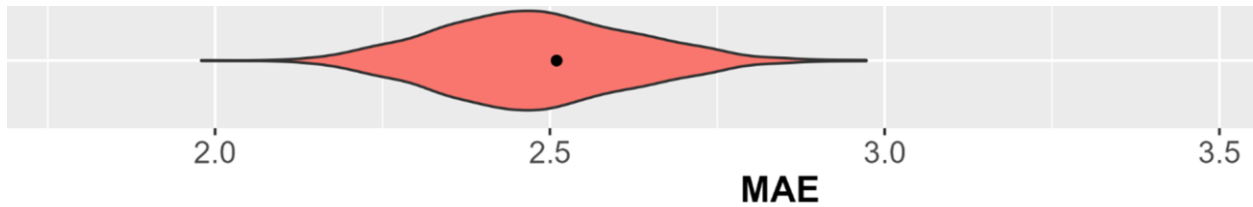
  	<u>Balance Penalty</u>	<u>Utility Prior</u>	<u>Utility Scaling</u>
Cell 3	No	0	
Cell 4	Yes	0	
Cell 5A	Yes	Based on N=50 from cell 4	100% of their original size
Cell 5B	Yes	Based on N=50 from cell 4	50% of their original size
Cell 5C	No	Based on N=50 from cell 4	100% of their original size

RESULTS

To measure the accuracy of the models built from each of the different design cells, we will explore the mean absolute error (MAE) of the models. Calculating the MAE involves assessing the average absolute difference between the predicted values from the hierarchical Bayesian model and the actual values in a set of test data, usually referred to as the holdout tasks. The larger the magnitude, the worse the model does at predicting the actuals—so the smaller MAE the better.

Historically, MAEs have been calculated using the point estimate from the HB model, which is typically calculated by taking the average value of all the draws. However, the whole point of using a Bayesian approach is to capture the uncertainty in the data. Thus, to calculate the MAE for this paper, we will leverage 1,000 draws from the HB model and create 1,000 MAEs. Then we will plot the distribution of the MAEs using a violin plot. For posterity’s sake, we will also plot the MAE based on the point estimate on the distribution chart. However, one will see in the following results the risks of only using point estimates to run analysis. In the example below (Figure 3.1), you can see the distribution of the 1,000 MAEs in red and the black dot on the chart represents the point estimate MAE.

Figure 3.1: Example MAE Distribution with Point Estimate MAE Included



Note—The point estimate MAE is calculated based on each individual's point estimate part-worth, which is the average of their posterior distribution. After finding the average of the posterior, we exponentiate each individual's point estimate, simulate the fixed task and report the average probability of choice and then report the difference from the stated frequencies. However, the point estimates can get distorted particularly when constraining price (a non-linear transformation) and the IIA property could also distort where the point estimate lies. Because of these two effects (1. order/sequence of averaging with non-linear transformation [before vs. after and within vs. across draws] and 2. IIA property of logit probabilities) the point estimate MAES (i.e., the black dots) are not required to be within the middle of the distribution and can even be found outside the distribution.

UNCONSTRAINED MODEL RESULTS

First modeling the data unconstrained, we can see that the distribution of MAEs for Cell 1 is further to the left in Figure 4.1 suggesting that it is the top performing design strategy when trying to predict responses from Cell 6. Figure 4.2 shows the likelihood that Cell 1 is better than the other cells (i.e., what percentage of the distribution of MAEs does not overlap with other cells). For example, Cell 1 is 87% likely to be better than Cell 2 and 100% likely to be better than Cell 3 and Cell 4 (i.e., there is no overlap in Cell 1's MAE distribution with Cell 3 and 4). Relative to Cell 1, it does appear that Cells 5b and 5c show promise.

Figure 4.1: Distribution of MAEs for Unconstrained Models when Predicting Cell 6

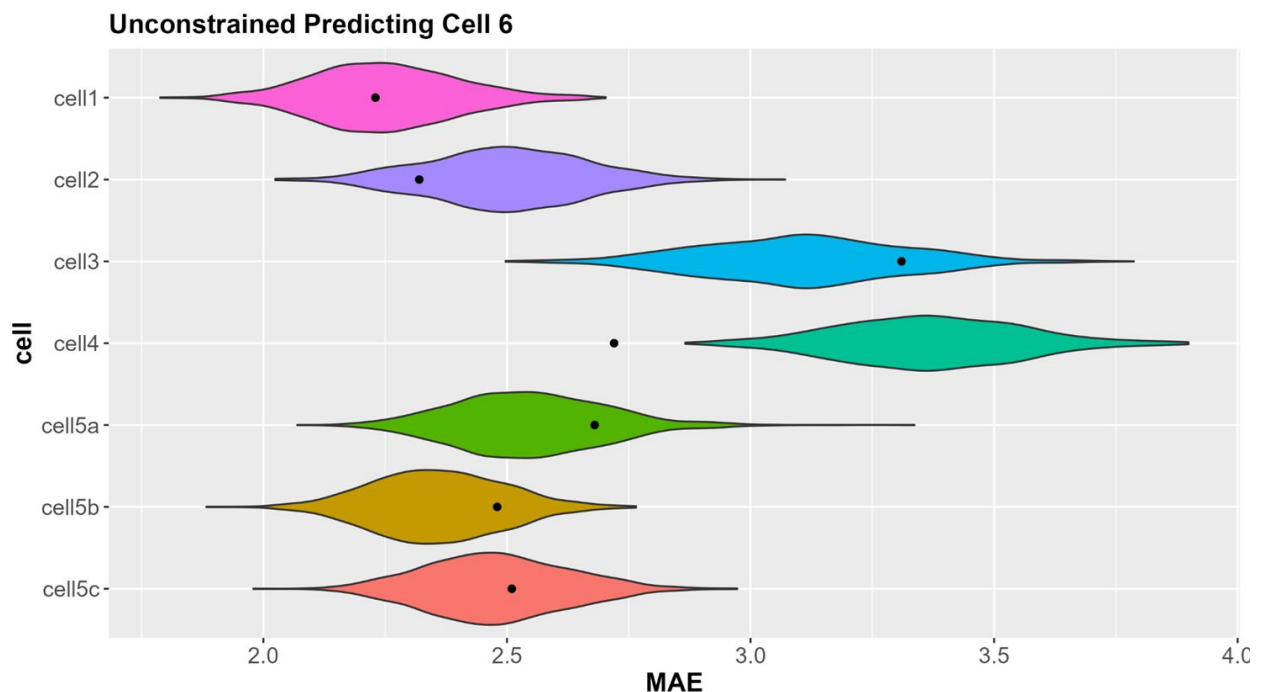


Figure 4.2: Likelihood of a Cell's MAE Outperforming the Other Cells when Predicting Cell 6

cell	Utility prior	Software	Type	cell1	cell2	cell3	cell4	cell5a	cell5b	cell5c
 1	zero	Sawtooth Lighthouse	Balanced and Overlap		87%	100%	100%	92%	71%	86%
 2	zero	Sawtooth Discover 2.0	Default	13%		99%	100%	58%	25%	46%
 3	zero	Julia	Min D-error	0%	1%		83%	2%	0%	1%
 4	zero	Julia	Level Balanced Min D-error	0%	0%	17%		0%	0%	0%
 5a	Utilities from 1st 50 Cell 4	Julia	Utility & Level Balanced Min D-error	8%	42%	98%	100%		18%	38%
 5b	Utilities from 1st 50 Cell 4	Julia	Utility & 50% Level Balanced Min D-error	29%	75%	100%	100%	82%		72%
 5c	Utilities from 1st 50 Cell 4	Julia	Utility & Min D-error	13%	54%	99%	100%	62%	28%	

It should also be noted that if one were to use the point estimate (black dot) instead of the draws, a researcher might come to very different conclusions. For example, they might claim that Cell 4 is significantly better than Cell 3—which, when looking at the draws, we know is not the case. Therefore, practitioners should remain cautious when drawing conclusions based only on the point estimate.

In addition to predicting Cell 6, we can take the data from Cell 1 and predict the two holdouts in Cell 2, 3, 4, 5a, 5b, and 5c. Then we can take the data from Cell 2 and predict the two holdouts in Cell 1, 3, 4, 5a, 5b, and 5c and so on.

Figure 4.3: Distribution of MAEs for Unconstrained Models when Predicting All Other Cells

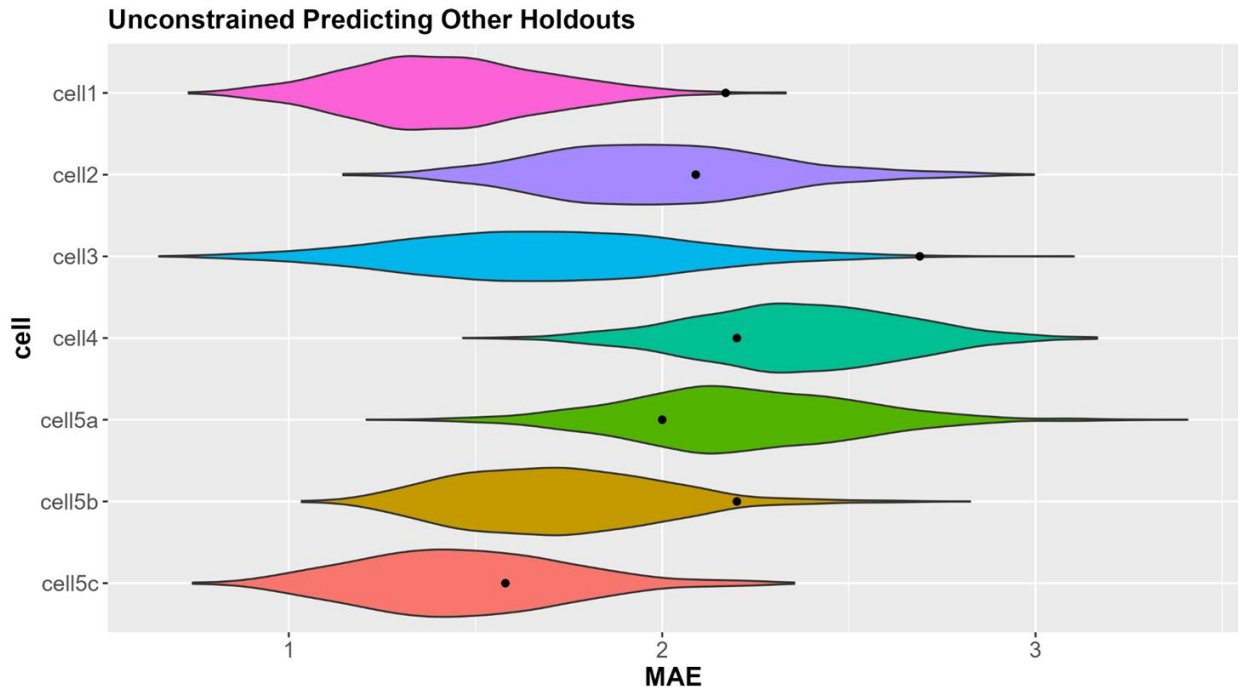


Figure 4.4: Likelihood of a Cell's MAE Outperforming the Other Cells when Predicting All Other Cells

cell	Utility prior	Software	Type	cell1	cell2	cell3	cell4	cell5a	cell5b	cell5c
 1	zero	Sawtooth Lighthouse	Balanced and Overlap		92%	72%	99%	97%	78%	55%
 2	zero	Sawtooth Discover 2.0	Default	8%		27%	81%	69%	25%	11%
 3	zero	Julia	Min D-error	28%	73%		92%	86%	52%	32%
 4	zero	Julia	Level Balanced Min D-error	1%	19%	8%		35%	5%	1%
 5a	Utilities from 1st 50 Cell 4	Julia	Utility & Level Balanced Min D-error	3%	31%	14%	65%		12%	4%
 5b	Utilities from 1st 50 Cell 4	Julia	Utility & 50% Level Balanced Min D-error	22%	75%	48%	95%	88%		27%
 5c	Utilities from 1st 50 Cell 4	Julia	Utility & Min D-error	45%	89%	68%	99%	96%	73%	

Figure 4.3 and 4.4 show Cell 1 continuing to outperform the other cells with Cell 4 being the worst at predicting out-of-sample data. (However, if only examining point estimates, one would conclude that Cell 5c is significantly better than Cell 1 when predicting the holdouts in all other cells).

CONSTRAINED MODEL RESULTS

Given that practitioners may use constraints in their model to avoid unrealistic utility estimates (i.e., high prices preferred to low prices vs. low prices preferred to high prices), we also wanted to explore the results when price is constrained to be negative. It is important to note that the application of constraints in a conjoint model should be carefully considered as constraints introduce assumptions or biases into the analysis. Constraints should align with the underlying business context and be based on informed judgments. Proper validation and sensitivity analysis should also be conducted to ensure that the imposed constraints do not overly restrict the model or compromise its predictive power.

Figure 5.1: Distribution of MAEs for Constrained Models when Predicting Cell 6

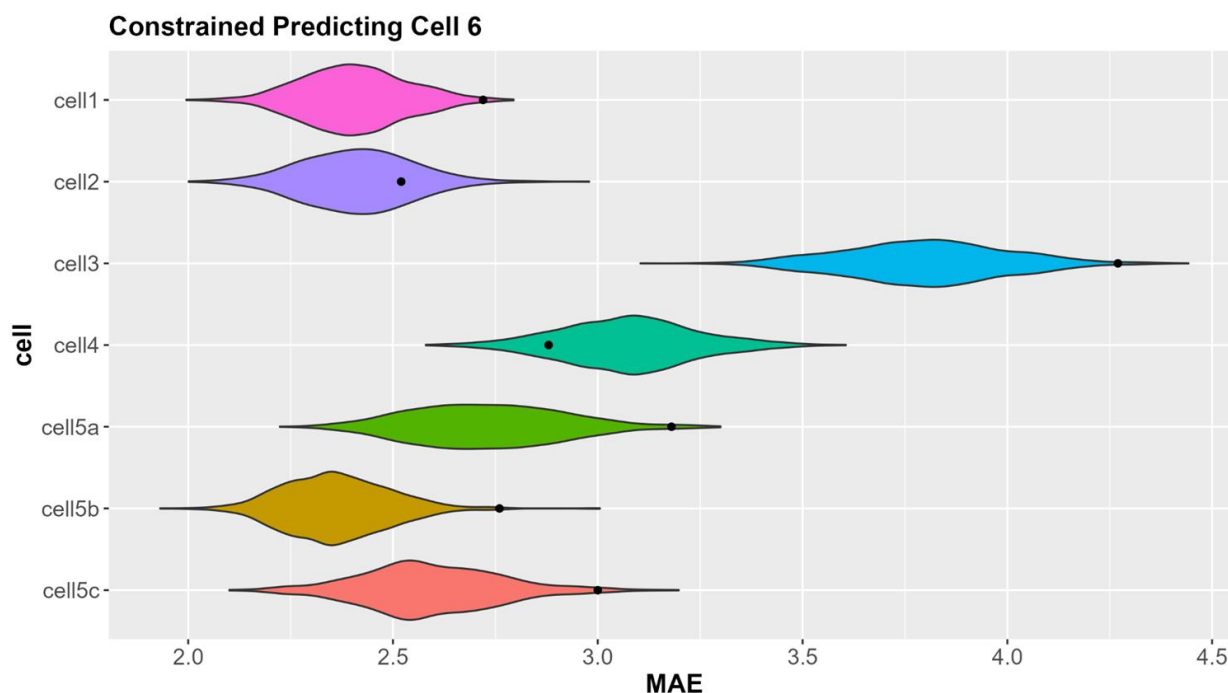


Figure 5.2: Likelihood of a Cell's MAE Outperforming the Other Cells when Predicting Cell 6

cell	Utility prior	Software	Type	cell1	cell2	cell3	cell4	cell5a	cell5b	cell5c
1	zero	Sawtooth Lighthouse	Balanced and Overlap		52%	100%	100%	93%	40%	81%
2	zero	Sawtooth Discover 2.0	Default	48%		100%	100%	92%	39%	80%
3	zero	Julia	Min D-error	0%	0%		0%	0%	0%	0%
4	zero	Julia	Level Balanced Min D-error	0%	0%	100%		9%	0%	3%
5a	Utilities from 1st 50 Cell 4	Julia	Utility & Level Balanced Min D-error	7%	8%	100%	91%		5%	29%
5b	Utilities from 1st 50 Cell 4	Julia	Utility & 50% Level Balanced Min D-error	60%	61%	100%	100%	95%		86%
5c	Utilities from 1st 50 Cell 4	Julia	Utility & Min D-error	19%	20%	100%	97%	71%	14%	

In the constrained models, Cells 1, 2, and 5b do very well with Cell 5c not far behind when predicting Cell 6 data (Figure 5.1). There is more overlap in the performance of Cells 1, 2 and 5b (Figure 5.2) suggesting that a utility balanced design is a viable option when constraints are needed.

Similar to the unconstrained model, when predicting all other cells combined, we see Cell 1 and Cell 5c as the best performers (Figure 5.3, 5.4). In all options, Cells 3 and 4 perform the worst but overall the MAE distributions are still relatively low (<4).

Figure 5.3: Distribution of MAEs for Constrained Models when Predicting All Other Cells

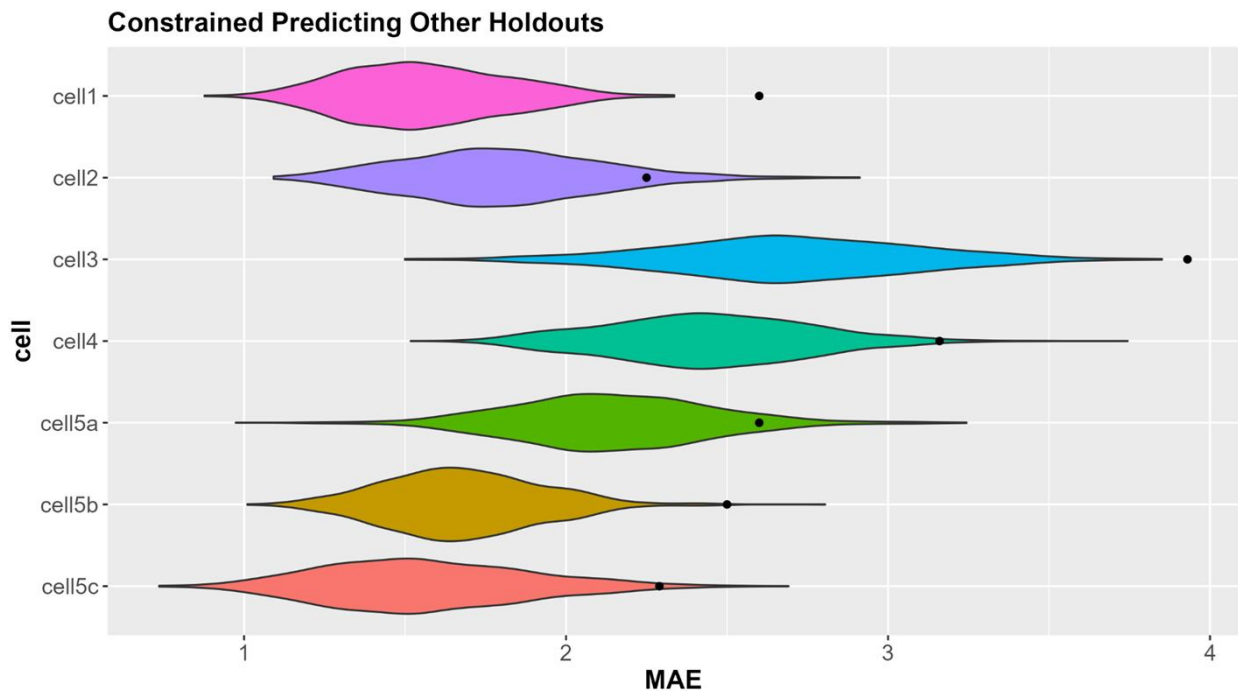


Figure 5.4: Likelihood of a Cell’s MAE Outperforming the Other Cells when Predicting All Other Cells

cell	Utility prior	Software	Type	cell1	cell2	cell3	cell4	cell5a	cell5b	cell5c
1	zero	Sawtooth Lighthouse	Balanced and Overlap		72%	99%	99%	93%	63%	49%
2	zero	Sawtooth Discover 2.0	Default	28%		96%	93%	79%	38%	29%
3	zero	Julia	Min D-error	1%	4%		30%	13%	2%	1%
4	zero	Julia	Level Balanced Min D-error	1%	7%	70%		25%	3%	3%
5a	Utilities from 1st 50 Cell 4	Julia	Utility & Level Balanced Min D-error	7%	21%	87%	75%		12%	10%
5b	Utilities from 1st 50 Cell 4	Julia	Utility & 50% Level Balanced Min D-error	37%	62%	98%	97%	88%		38%
5c	Utilities from 1st 50 Cell 4	Julia	Utility & Min D-error	51%	71%	99%	97%	90%	62%	

RESPONDENT REACTIONS TO THE DIFFERENT CELLS

From a model standpoint, utility balanced designs, particularly Cells 5b and 5c, seem to be viable options when creating designs that can perform on par with Sawtooth Software designs for this dataset. Next, we want to address how respondents might feel about the different designs.

Using a 5-point semantic differential, we asked respondents to rate the experiment on different dimensions (i.e., long vs. short, difficult vs. easy). Overall, all cells were easy, enjoyable, and appealing (Figure 6.1). Double-clicking into the “easy vs. difficult,” we see that over two-thirds of respondents classified the exercise as “easy” regardless of what cell they were in (Figure 6.2). Therefore, one could conclude that for this survey, for these respondents, a utility balanced design is no more difficult than a traditional design.

In Figure 6.3, we can see some additional metrics such as length of interview (LOI) and drop-off rate. LOI does appear to be higher for the utility balanced cells as respondents no longer have tasks that are “no brainers”—but that does not seem to impact respondent opinion, error, or drop-off rates.

Figure 6.1: Respondent Ratings (Means) of Cell Experience

cell	Utility prior	Software	Type	Long vs. Short	Difficult vs. Easy	Unappealing vs. Appealing	Dull vs. Fun	Unenjoyable vs. Enjoyable
1	zero	Sawtooth Lighthouse	Balanced and Overlap	3.4	4.4	4.2	3.9	4.2
2	zero	Sawtooth Discover 2.0	Default	3.5	4.4	4.3	4.1	4.3
3	zero	Julia	Min D-error	3.5	4.4	4.2	4.1	4.3
4	zero	Julia	Level Balanced Min D-error	3.5	4.4	4.2	4.0	4.3
5a	Utilities from 1st 50 Cell 4	Julia	Utility & Level Balanced Min D-error	3.5	4.4	4.2	4.0	4.2
5b	Utilities from 1st 50 Cell 4	Julia	Utility & 50% Level Balanced Min D-error	3.5	4.4	4.2	4.1	4.3
5c	Utilities from 1st 50 Cell 4	Julia	Utility & Min D-error	3.4	4.4	4.2	4.0	4.1
6	zero	Sawtooth Lighthouse	Balanced and Overlap (1 version)	3.5	4.4	4.1	4.0	4.2

Figure 6.2: Frequencies of Easy versus Difficult by Cell

cell	Utility prior	Software	Type	Easy ← → Difficult				
				1	2	3	4	5
1	zero	Sawtooth Lighthouse	Balanced and Overlap	68%	12%	14%	2%	5%
2	zero	Sawtooth Discover 2.0	Default	69%	12%	13%	2%	4%
3	zero	Julia	Min D-error	66%	18%	12%	1%	3%
4	zero	Julia	Level Balanced Min D-error	68%	12%	13%	3%	5%
5a	Utilities from 1st 50 Cell 4	Julia	Utility & Level Balanced Min D-error	67%	15%	12%	2%	5%
5b	Utilities from 1st 50 Cell 4	Julia	Utility & 50% Level Balanced Min D-error	68%	13%	13%	2%	4%
5c	Utilities from 1st 50 Cell 4	Julia	Utility & Min D-error	67%	12%	13%	3%	4%
6	zero	Sawtooth Lighthouse	Balanced and Overlap (1 version)	68%	12%	12%	2%	5%

Figure 6.3: Additional Metrics Captured by Cell

cell	Utility prior	Software	Type	Median LOI mins	% LT 10 min	% GT 10 min	Drop-off Rate
1	zero	Sawtooth Lighthouse	Balanced and Overlap	9.18	57%	43%	4.4%
2	zero	Sawtooth Discover 2.0	Default	9.95	51%	49%	3.7%
3	zero	Julia	Min D-error	9.55	54%	46%	5.2%
4	zero	Julia	Level Balanced Min D-error	9.07	58%	42%	3.6%
5a	Utilities from 1st 50 Cell 4	Julia	Utility & Level Balanced Min D-error	10.28	48%	52%	3.0%
5b	Utilities from 1st 50 Cell 4	Julia	Utility & 50% Level Balanced Min D-error	9.77	52%	48%	4.0%
5c	Utilities from 1st 50 Cell 4	Julia	Utility & Min D-error	10.08	50%	50%	4.1%
6	zero	Sawtooth Lighthouse	Balanced and Overlap (1 version)	9.72	53%	47%	6.2%

COMPARING PREDICTIONS OF UTILITY VS. NON-UTILITY BALANCED DESIGNS

One additional finding to be discussed is the ability of the model from a utility balanced design to predict responses from non-utility balanced design cell. In Figure 7.1 we can see that Cells 5a, 5b and 5c have low MAEs when predicting other utility balanced cells and non-utility balanced cells (Cell 1, 2, 3, and 4). However, the non-utility balanced cells (Cell 1, 2, 3, and 4) struggle to predict utility balanced cells (5b and 5b).

Figure 7.1: Average MAEs per Cell when Predicting Other Cells

cell	Utility prior	Software	Type	cell1	cell2	cell3	cell4	cell5a	cell5b	cell5c	cell6	Average	insample
1	zero	Sawtooth Lighthouse	Balanced and Overlap		2.06	1.49	2.14	4.10	2.68	3.00	2.44	2.56	2.06
2	zero	Sawtooth Discover 2.0	Default	2.41		2.20	1.97	3.16	3.05	2.83	2.62	2.60	1.69
3	zero	Julia	Min D-error	2.53	2.46		2.52	4.58	3.31	3.33	3.10	3.12	1.55
4	zero	Julia	Level Balanced Min D-error	2.73	2.07	2.02		3.48	2.67	3.12	2.79	2.70	1.70
5a	Utilities from 1st 50 Cell 4	Julia	Utility & Level Balanced Min D-error	2.64	2.54	2.25	2.35		2.21	1.81	2.99	2.40	2.71
5b	Utilities from 1st 50 Cell 4	Julia	Utility & 50% Level Balanced Min D-error	2.29	2.68	2.14	2.43	3.55		2.74	2.21	2.58	2.27
5c	Utilities from 1st 50 Cell 4	Julia	Utility & Min D-error	2.12	2.38	1.73	1.95	3.17	1.87		2.01	2.18	2.06

This makes us wonder if people are responding to the choice exercises differently, relative to which cell they are in. One hypothesis is that by using a utility balanced design, we might be priming people to answer the fixed tasks differently than they would if it were a standard design. Initial exploration seems to suggest that the utility balanced cells are potentially using the none alternative differently.

If we look at a different study and compare a Lighthouse Studio design (Cell 1) to a Julia, utility balanced design, we can see that when the none is excluded (Figure 7.2), the Julia cells (JL) perform much better than the Lighthouse Studio cells (LH). But, when we include the none in the model (Figure 7.3), the Lighthouse Studio design cells do much better than the Julia cells.

Figure 7.2: Comparing MAEs of a Lighthouse Design versus a Julia Utility Balanced Design, Excluding the None Option

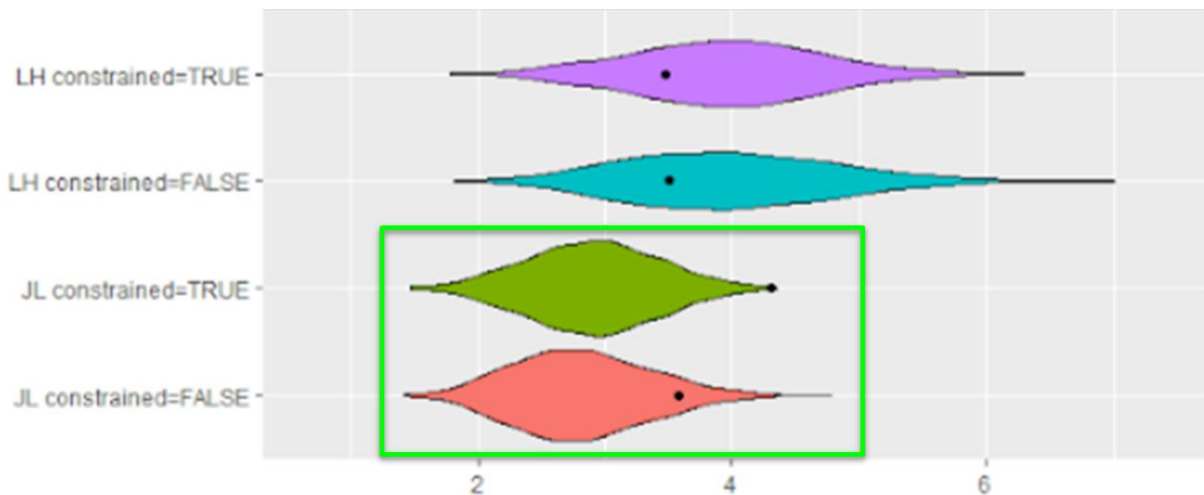
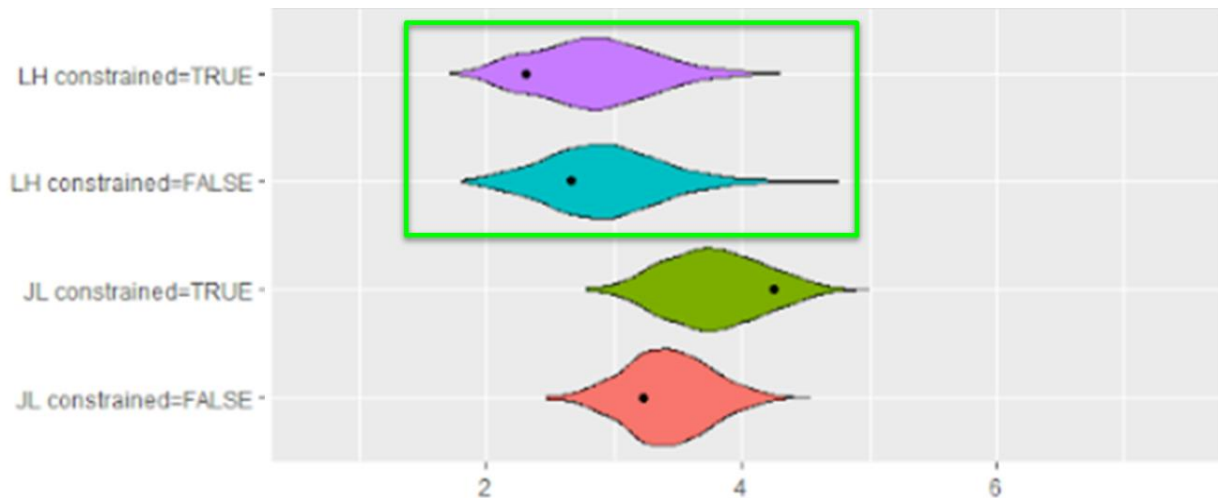


Figure 7.3: Comparing MAEs of a Lighthouse Design versus a Julia Utility Balanced Design, Including the None Option



To explore the potential influence of the design alternatives on respondent behavior, we examined the willingness-to-pay (WTP) values for each design. We hypothesized that if the design was influencing respondent behavior, we might see differences in WTP values between the designs. Our results showed that for one product, all four designs produced similar WTP estimates. However, for another product, there were significant differences between the Lighthouse and Julia designs. This finding suggests that further research is necessary to uncover the factors that may be driving these differences.

CONCLUSION

In this study, we investigated the use of utility balance designs in CBC experiments. Our results suggest that leveraging estimated prior utilities to inform the CBC design can be valuable, particularly if the researcher plans to constrain the model. The authors encourage other practitioners to test out this approach by exploring the literature around utility balanced and/or two-stage designs (additional references can be found in References) as well as reaching out should you choose to leverage Numerious' Julia designer package.

However, further research is necessary to explore the potential influence of the design on respondent behavior and to uncover the factors that may be driving differences in WTP estimates between designs. In addition, caution should be exercised when using utility balance designs, as they may result in sparse data at the interaction level. In these cases, a standard balance and overlap design or an alternative specific design may be more appropriate.

NEXT STEPS

If you are curious to try out a utility balanced design, reach out to the authors of this paper. We've created a private GitHub repo and welcome anyone that wants access upon request. We've only just scratched the surface on features and would love to build a more robust, open source tool together.



Megan Peitz



Trevor Olsen

REFERENCES

- Bliemer, Michiel C.J. & John M. Rose (2011) “Experimental design influences on stated choice outputs: An empirical study in air travel choice,” *Transportation Research Part A: Policy and Practice*, Volume 45, Issue 1, Pages 63–79. ISSN 0965-8564, <https://doi.org/10.1016/j.tra.2010.09.003>.
- Huber, Joel and Klaus B. Zwerina (1996) “The Importance of Utility Balance in Efficient Choice Designs,” *Journal of Marketing Research* 33 (August), 307–17.
- Kuhfeld, Warren, (1997) “Efficient Experimental Designs Using Computerized Searches,” *Sawtooth Software Conference Proceedings*, 71–86.
- McFadden, Daniel (1974), “Conditional Logit Analysis of Qualitative Choice Behavior,” in *Frontiers of Econometrics*, P. Zarembka, ed. New York: Academic Press, 105–42.
- Rose, John M. and Michiel C. J. Bliemer (2009) “Constructing Efficient Stated Choice Experimental Designs,” *Transport Reviews*, 29:5, 587–617, DOI: 10.1080/01441640902827623.
- Rossi, P. E., Allenby, G. M., and McCulloch, R. (2005). *Bayesian Statistics and Marketing*. Wiley-Blackwell.
- Sawtooth Software (2018) “Discover-CBC: How and Why It Differs from Lighthouse Studio’s CBC Software,” <https://content.sawtoothsoftware.com/assets/0684c377-88bf-4478-8535-e3841079afba>
- Traets, F., Sanchez, D.G. and Vandebroek, M. (2020) “Generating Optimal Designs for Discrete Choice Experiments in R: The idfix Package,” *Journal of Statistical Software*, 96, 3 (Nov. 2020), 1–41. DOI: <https://doi.org/10.18637/jss.v096.i03>.
- Walker, J.L., Y. Wang, M. Thorhauge and M. Ben-Akiva (2018) “D-efficient or Deficient? A Robustness Analysis of Stated Choice Experimental Designs,” *Theory and Decision*, 84: 215–238.

COMMENTS ON “BUILDING DESIGNS FOR INDIVIDUAL-LEVEL ESTIMATION”

KEITH CHRZAN

SAWTOOTH SOFTWARE

First, I want to say that I think Megan and Trevor may have under-sold what they did in this paper. From their presentation, I had assumed that their Julia designs replicated the Huber and Zwerina (1996) suggestions (a) to modify an orthogonal or level-balanced design to promote utility balance, and (b) to base their utility priors on utilities from pre-test respondents. They actually did more than that. Not only did they use the prior estimates of β to make efficient designs in Julia, they also used the distribution of β s from their pretest respondents to make a Bayesian efficient design, as suggested by Sándor and Wedel (2002). That’s a neat trick that might be new to some Sawtooth Software users. Providing a free tool that makes these designs was a nice bonus.

That said, there’s also evidence in the literature that optimizing design based on prior utilities may not produce robust utility estimates. If respondents’ utilities deviate much from the priors, the resulting models may become worse than would have resulted from a simpler orthogonal (or even random) design (Walker et al. 2018). Indeed, the Balanced Overlap design that is Cell 1 in Megan and Trevor’s experiment bests the optimal Julia designs in Cells 3 to 5c in many of the tests reported in the paper. Perhaps this isn’t too surprising, given that the Balanced Overlap strategy, by design, seeks to be robust across many conditions.

One thing struck me as a little odd, at least based on some internal testing we did as part of developing our Discover product. Whereas we found the Discover design to perform slightly better than the Balanced Overlap design strategy that is the default in our Lighthouse Studio software, Megan and Trevor found the reverse in their MAE comparisons. It’s not clear why the Balanced Overlap cell should have done better than the Discover cell in Megan and Trevor’s tests: both involve uninformative priors, both featured essentially the same amount of level overlap, and the Discover designs feature better one-way and two-way frequency balance, especially when considering the within-version balance.

Using empirical utilities from Megan and Trevor’s study I programmed robotic respondents to complete both the Balanced Overlap and the Discover designs Megan and Trevor’s used. In terms of parameter recovery, I found a very small difference favoring the Discover design: utilities from respondents answering the Discover design question were correlated at 0.948 with the true utilities versus 0.942 for the utilities from the Balanced Overlap design, roughly a 10% reduction in error. The difference in findings between this parameter recovery experiment and the MAE comparison in the paper is small and could easily result from the particular random split between test and holdout respondents in Megan and Trevor’s study.



REFERENCES

- Huber, J. and K.B. Zwerina (1996) “The importance of utility balance in efficient choice designs,” *Journal of Marketing Research*, **33** (August), 307–17.
- Sándor, Z. and M. Wedel (2002) “Profile construction in experimental choice designs for mixed logit models,” *Marketing Science*, **21** (4), 455–475.
- Walker, J.L., Y. Wang, M. Thorhauge and M. Ben-Akiva (2018) “D-efficient or deficient? A robustness analysis of stated choice experimental designs,” *Theory and Decision*, **84**: 215–238.

HOW MANY ITERATIONS DO WE NEED?

GUIDELINES FOR THE RIGHT NUMBER OF BURN-IN AND USED DRAWS IN HIERARCHICAL BAYES ESTIMATION

PETER KURZ

MAXIMILIAN RAUSCH

BMS - MARKETING RESEARCH + STRATEGY

INTRODUCTION

Conjoint analysis is widely used for preference estimation, with the Hierarchical Bayes Multinomial Logit (HB-MNL) method being popular for individual-level analysis. When discussing HB estimation settings, researchers often debate the appropriate number of iterations. Determining the required number of burn-in iterations, saved draws, and thinning factor is crucial for obtaining reliable part-worth utilities. Additionally, choosing between using HB draws for simulation or calculating point estimates adds complexity. The correct thinning factor helps avoid serial correlation and Bayesian error caused by oscillations that can influence results. Should one use each 10th, each 50th, each 100th or even larger thinning, to avoid oscillations that influence the results? This paper aims to explain the significance of these factors in obtaining reliable part-worth utilities and provide guidelines for everyday HB estimation usage.

BACKGROUND

The initial version of CBC/HB by Sawtooth Software, delivered on a floppy disk in 1999, had default settings of 1,000 burn-in iterations and 1,000 saved draws. The current version (5.7) uses 10,000 as the default for both burn-in and saved draws and aggregates the point-estimates over the 10,000 saved draws with a thinning factor of 10. Due to advancements in computational power, practitioners often opt for higher iteration numbers, such as 50,000/50,000, 100,000/100,000, or even 190,000/10,000. Thinning is not very often discussed and most of the studies use a factor of 10. These examples are based on practitioners' extensive experience from numerous studies conducted over the years, although they lack academic confirmation. What are the observed reasons behind these different findings? A long burn-in phase should guarantee convergence of the model—which is necessary to get stable part-worth utilities. A large number of used draws and thinning the draws should guarantee that long-term oscillations and reasons for poor mixing cancel out. Sparse data (lots of parameters and limited numbers of choice tasks) needs more draws for the Markov Chain to reach stationarity and for long-term effects (Bayesian error) to be more likely to disappear.

This paper aims to explore these issues and investigate the effects of burn-in, used draws, and thinning factor. Therefore, an intensive analysis of Markov Chain behavior is needed to figure out if models reach stationarity and what behavior the chains have and give researchers guidelines for everyday use of HB. Furthermore, we want to give hints as to how long the models need to run until stationarity is reached and allow forecasts of run lengths in early stages of the study setup. This can help researchers plan the needed time for analytics, sometimes even before fieldwork starts.

Monte Carlo Simulation Study

To evaluate models with different complexities and sparseness, we conducted a Monte Carlo simulation study with 13,824 datasets. The experimental factors included Sample Size, # of Attributes, # of Levels, # of Tasks, Concepts per Task, # of burn-in draws, # of used draws, Thinning factor, and heterogeneity versus homogeneity in the data (Table 1).

Table 1

Sample Size	500	1,000	2,000	
Number of Attributes	6	12	18	24
Number of Levels	18	36	72	144
# of Tasks per Respondent	10	15	20	
# of Concepts per Task	8 for all simulated models			
Number of burn-in draws	1,000	10,000	50,000	100,000
Number of used draws	1,000	10,000	50,000	100,000
Thinning-Factor	1	10	100	
Heterogeneity	homogeneous	heterogeneous		

Simulation Guidelines

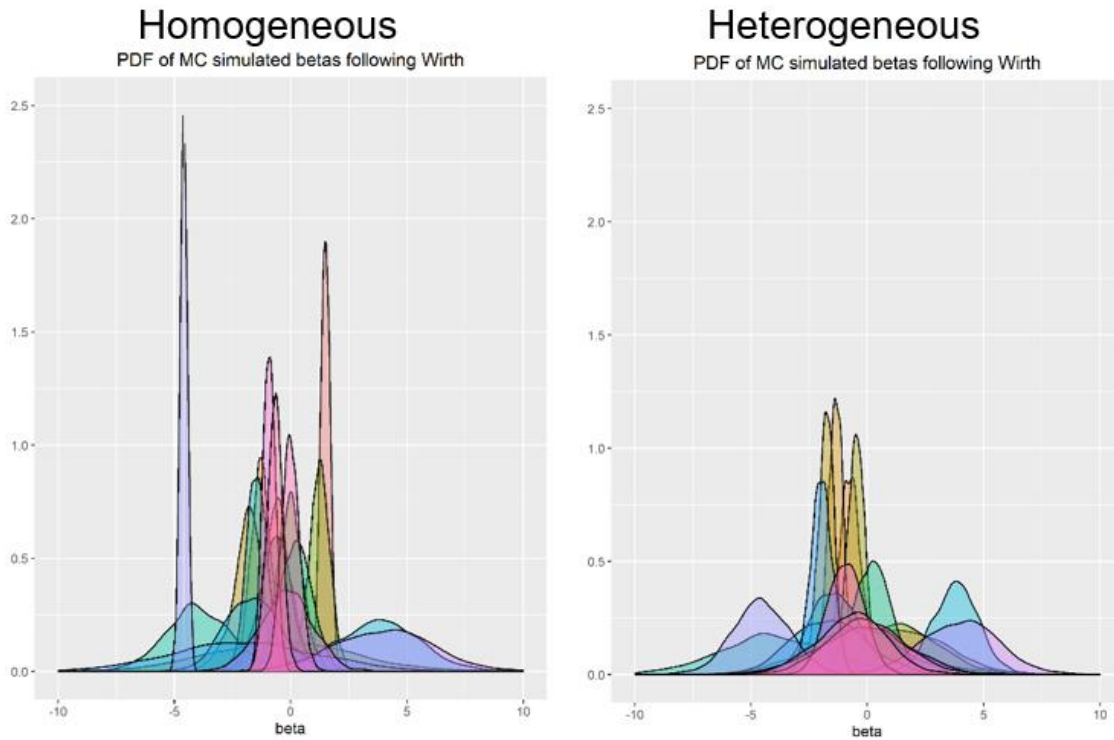
Our simulation guidelines are based on Ralf Wirth's book (2010a). For each treatment, individual part-worths were generated following a mixture of normal distributions. The elements of the initial true part-worth utilities followed a uniform distribution ranging from -5 to 5, which aligns with the ranges commonly observed in empirical datasets. Wirth analyzed studies collected from GfK, and we corroborated these findings by analyzing over 900 datasets from Kantar.

To capture preference heterogeneity, we generated the covariance matrix using a combination of gamma and uniform draws. Random draws were generated from a gamma distribution with shape parameter 0.7 and scale parameter 2.0. Since the gamma distribution is highly skewed with large parts of its mass near zero, additional random draws were generated from a uniform distribution $U(0.2,2)$ and added to the draws from the gamma distribution in order to avoid variances that are too small (Hein, Kurz, Steiner 2020).¹ Including both homogeneity and heterogeneity in the data generation allowed us to examine the influence of heterogeneity on the number of iterations required for the Markov Chain to reach stationarity.

The plots in Figure 1 illustrate the differences between homogeneous and heterogeneous distributions based on one of our smaller models. The left plot displays normal distributions with higher peaks and fewer draws in the tails, indicating a homogeneous distribution. In contrast, the right plot shows more draws in the tails, indicating a more heterogeneous distribution.

¹ The setting of the gamma distribution and the uniform distribution to 2 result in the later used factor 2 for calculating draws for heterogeneous samples.

Figure 1



This specification is used later as factor of heterogeneity in the sparseness calculation.

EXPERIMENTAL DESIGNS

To assess the quality of our models, we generated separate test and validation samples. The test sample was used for computing the HB models, while the validation sample was utilized to measure reproducibility rates and various goodness-of-fit measures. To achieve optimal experimental conditions, we employed different designs for the two samples. We ensured that the design error was minimal and did not unduly influence the results by testing the designs for D-efficiency and only accepting designs with a D-efficiency score not less than 95%.

In simulation studies, it is feasible to create two distinct datasets for estimation and validation, which is considered the gold standard. In empirical studies, achieving this requires doubling the costs. But, in our simulation study, generating separate synthetic respondents for estimation and validation incurred no additional costs.

GOODNESS-OF-FIT MEASURES WE USE IN THIS PAPER

To examine the influence of the number of iterations on the quality of the resulting part-worth utilities and the accuracy of recovering the real answers, we used the following goodness-of-fit measures:

- Root-Likelihood (RLH) of the estimated models
- Root Mean Square Error (RMSE) of the simulated preference shares compared to the real shares
- Hit Rate between the known true answers and simulated answers

The rigorous test conditions, involving the use of both validation and estimation samples, resulted in lower hit rates and higher RMSE values than in-sample testing would yield. However, our approach produces results closer to what would be achieved with real data collected in real online panels.

DATA COMPUTATION

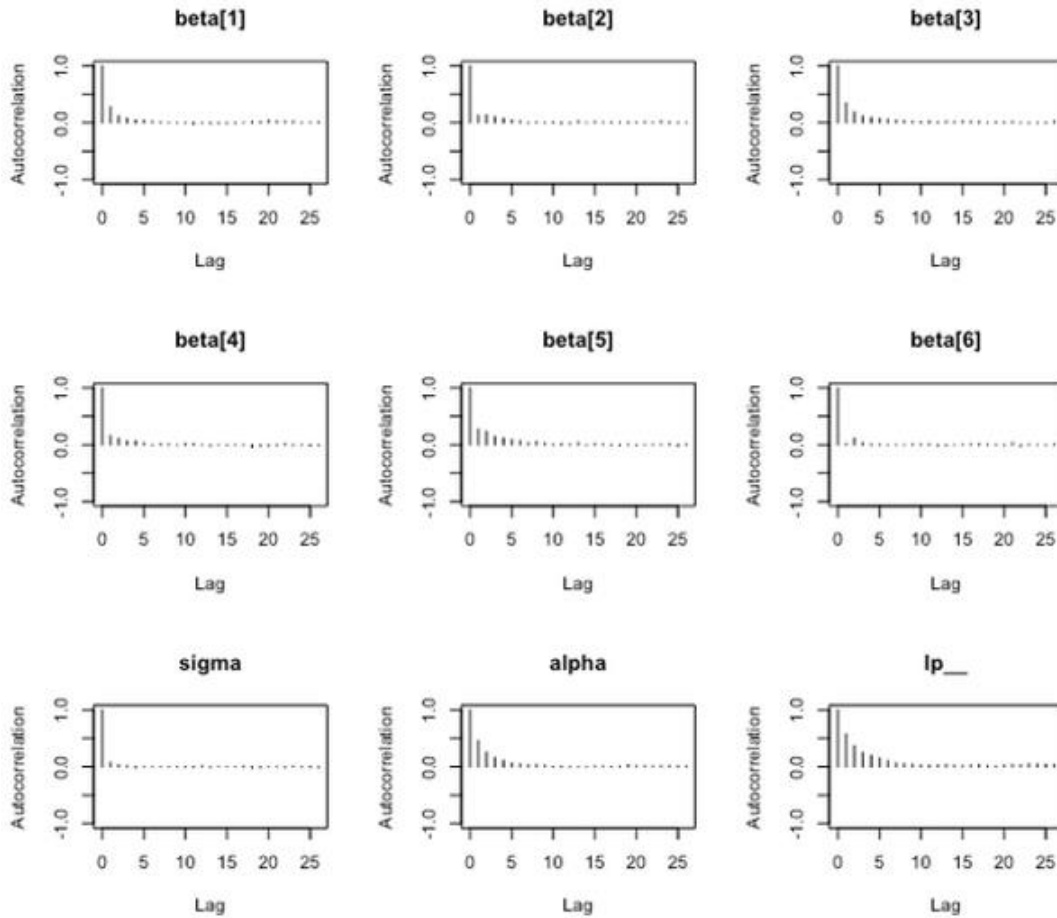
To compute the large number of simulated datasets (13,824), we utilized the R “bayesm” package. This choice allowed us to script the process and run it on multi-core computers. Additionally, due to the high number of estimated draws and the large dimensions of the vectors and matrices involved, we had to modify parts of the R and C++ code to avoid memory and allocation problems. The final execution of the scripts was performed on a UNIX-Exascale Supercomputer to obtain results within an acceptable timeframe. However, it is worth noting that the computations can be reproduced using CBC/HB, as we utilized the same standard settings as the Sawtooth software. The only differences were the variations in burn-in draws and saved draws, which can easily be adjusted in CBC/HB. We chose not to use Sawtooth Software’s CBC/HB due to the time constraints associated with running 13,824 models.

TESTING THE MARKOV CHAINS FOR STATIONARITY

To test our Markov Chains for stationarity, we typically employ the convergence diagnostics procedures from the R package “coda.” The procedures we use include tests for autocorrelation, the Gelman and Rubin diagnostic, and the Raftery and Lewis diagnostic.

The plot in Figure 2 illustrates the autocorrelation diagnostic for different parameters. For example, `beta[2]` demonstrates a well-estimated parameter, with fast convergence and autocorrelation disappearing quickly with a small lag factor. In contrast, `lp__` exhibits slow convergence and requires a larger lag factor until autocorrelation disappears. If such parameters are present in the model, it is advisable to rerun the model with more iterations and increase the burn-in phase.

Figure 2

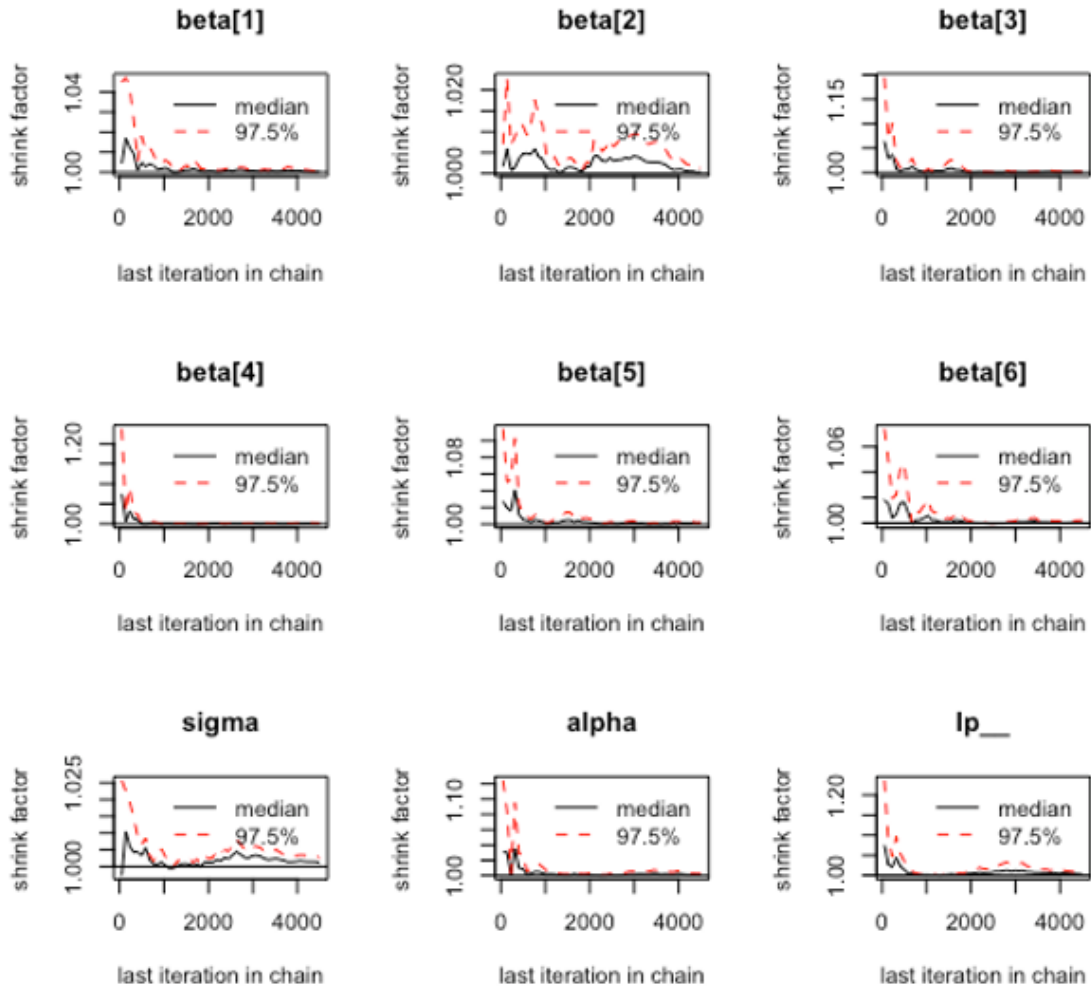


The Gelman and Rubin Potential Scale Reduction Factor is a frequently used test to assess whether a Markov Chain Monte Carlo (MCMC) model, begun from different starting points, has converged to the same location. In this illustration, we used 10 different starting points, ran 8,000 iterations, discarded the first 4,000 as burn-in, and used the next 4,000 to calculate the Potential Scale Reduction Factor (PSRF). The Gelman and Rubin approach involves decomposing the variance of all 10 chains into within-chain and between-chain variances and checking for significant differences between the chains.

As is standard practice, we aggregated the calculated parameter for each respondent, resulting in one PSRF parameter for each estimated parameter of the model. In most of the literature, convergence is accepted when the PSRF is less than 1.1.

Figure 3 shows the PSRFs for all parameters as the chain grows to the full 4,000 iterations. For **beta[2]**, neither the median nor the 97.5% interval is particularly stable. However, by around 3,000 iterations, the PSRF approaches 1.01, indicating that all 10 chains converge to the same point.

Figure 3



For “ $lp_$,” we see that this parameter gets closer to 1.0, but it still exhibits more autocorrelation. Therefore, in light of the Gelman and Rubin diagnostics, we would recommend running more iterations and using a longer burn-in phase to ensure that stationarity is reached. Since we are unable to report all parameters from our 13,482 models (which have between 18 and 138 parameters each), we used an aggregated measure that we called A-PSRF. To calculate A-PSRF, we summed up all PSRF values from each of our models and divided them by the number of parameters of the respective model. This aggregation step resulted in one A-PSRF value for each model. If the A-PSRF is smaller than 1.10, the model can be assumed to have converged.

SPARSENESS INDEX

In addition to the A-PSRF measure, we decided to calculate a sparseness index that employs various model parameters to assess the data density of the model. The amount of information captured by a model relies on several factors, including the number of respondents, the number of tasks presented to each respondent, and the number of concepts per task. These factors are counterbalanced by the number of attributes and levels in the model, which corresponds to the

number of parameters that need to be estimated. In order to facilitate the assessment of a model and its complexity, we devised a sparseness index that takes into account both the information collected and the need for information based on number of parameters.



Sparseness Index

to get an idea on how many iterations are needed

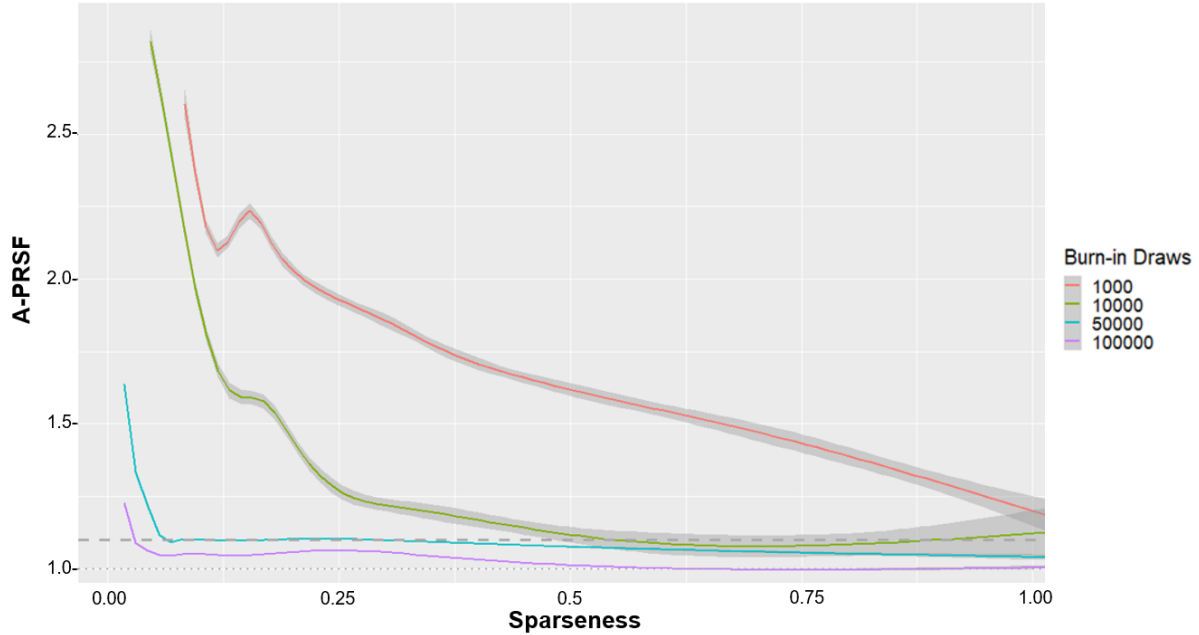
$$Sparseness = nConc \times \frac{\ln(N)}{(nLevels - nAtt) \times het} \times \frac{nTasks}{100}$$

<u>nConc</u>	number of concepts per task (always 8 in our simulated data)
<u>nLevels</u>	number of levels
<u>nTasks</u>	number of tasks per respondent
N	sample size
het	set to 1: for homogeneous set to 2: random coefficients from gamma and uniform distribution are used to simulate heterogeneity

The sparseness index² diminishes when less information is collected from respondents, and when the parameters to estimate become more numerous. This index can be utilized to illustrate the relationship between the complexity of the models and the A-PSRF convergence measures for our various models (see Figure 4).

² In empirical studies it is not that easy to set the correct factor for heterogeneity, because one doesn't know how homogeneous or heterogeneous the sample is. Factor 2 is a good approximation, based on more than 1,200 empirical studies that are used to find the correct setting for our simulation study.

Figure 4

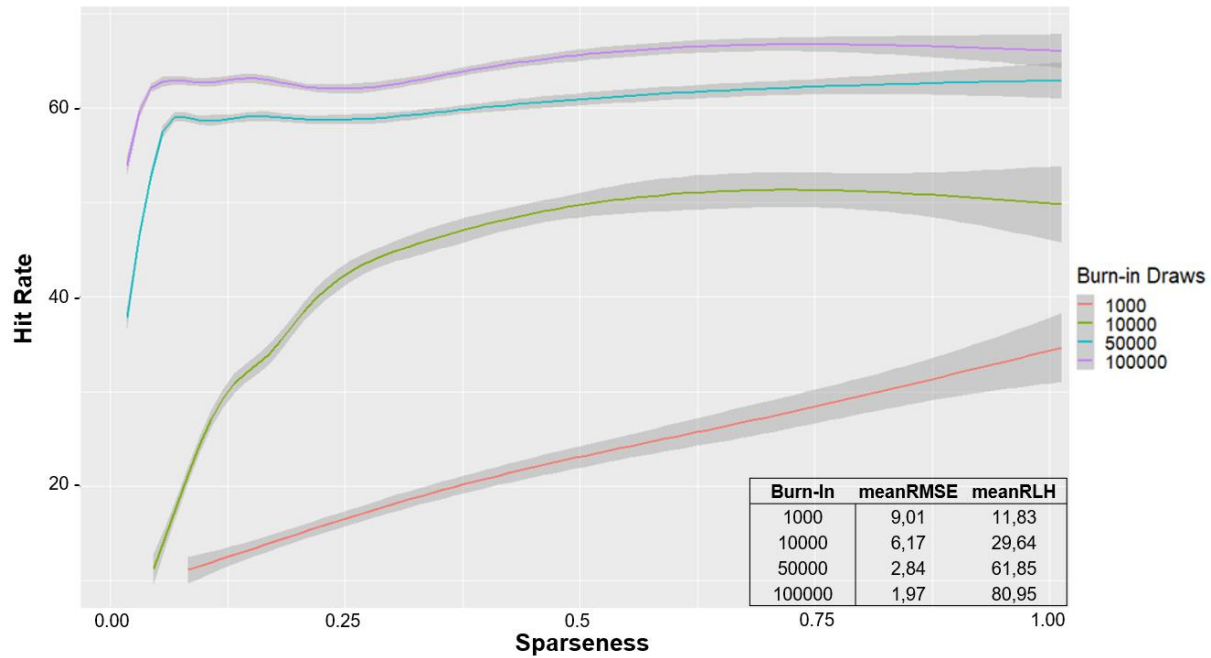


Sparseness and Hit-Rate (**Smoothing line**) plotted for all 13,824 models.

Figure 4 clearly demonstrates the dependence between burn-in draws, the A-PSRF measure, and the sparseness index. By adopting a convergence threshold of 1.1, the customary standard, we observe that increasing the number of burn-in draws brings us closer to convergence for larger models. For any of the tested models, 1,000 burn-in draws are insufficient to achieve convergence. However, for a sparseness index greater than approximately 0.5, 10,000 burn-in draws may be adequate for convergence. Additionally, for 50,000 and 100,000 burn-in draws, we notice that convergence can be attained even with relatively low values of the sparseness index. Notably, 100,000 draws consistently perform slightly better than 50,000 draws.

Figure 5 shows that the out-of-sample hit rate, often employed as a measure of goodness of fit for conjoint studies, exhibits a similar trend. It is evident that with only 1,000 burn-in draws, the hit rate remains relatively low even with substantial amounts of information (high sparseness index). However, 10,000 draws can yield satisfactory results when there is sufficient information in the data. Furthermore, for 50,000/100,000 burn-in draws, we achieve quite favorable hit rates even with lower sparseness indices. Once again, a larger number of burn-in draws leads to even higher hit rates.

Figure 5



Sparseness and Hit-Rate (**Smoothing line**) plotted for all 13,824 models.

RAFTERY AND LEWIS DIAGNOSTIC

A second measure often used to test convergence is the Raftery and Lewis diagnostic. The R&L diagnostic is popular because it is a one-chain diagnostic, so one need only compute one long initial chain. (Remember that the Gelman and Rubin diagnostic needs at least 2 chains; more are even better.) A one-chain diagnostic is easier and saves computational time. The Raftery and Lewis diagnostic indicates whether the Markov chain has reached stationarity and delivers parameters one needs for running the MCMCs.

R&L give the following values:

1. The minimum number of iterations that should be run
2. A suggested number of burn-in iterations
3. The recommended thinning interval (keep every k^{th} draw)

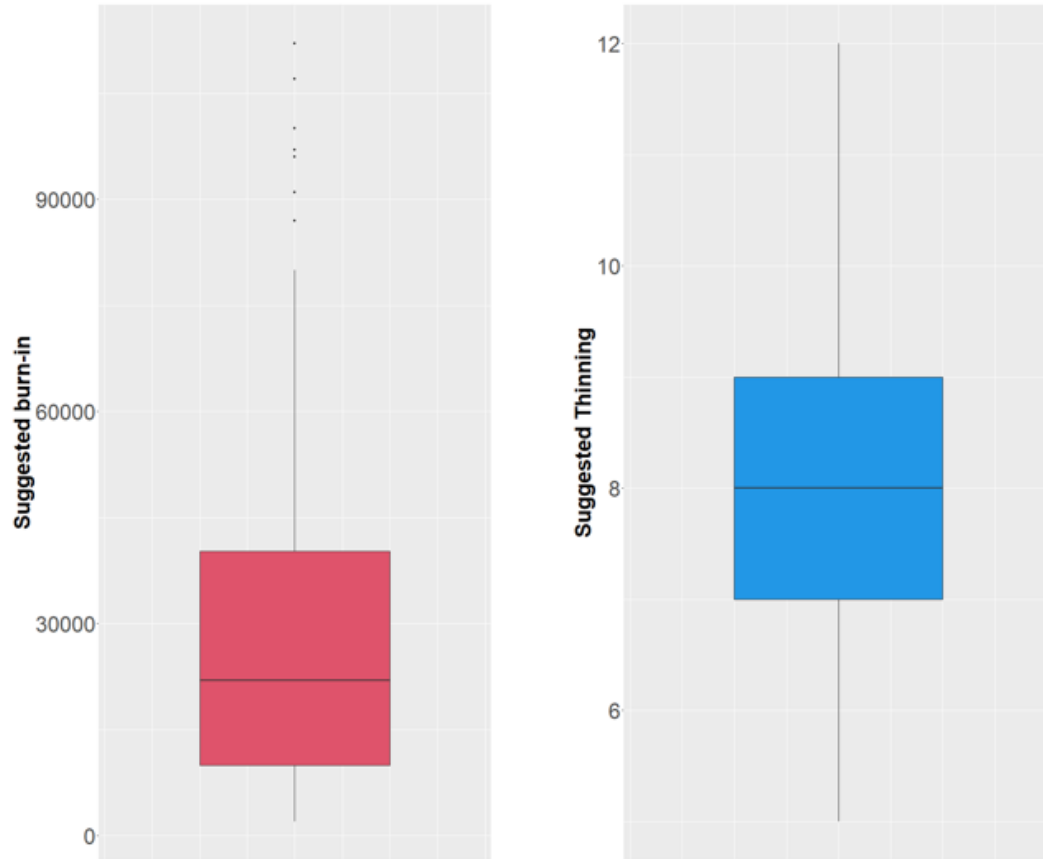
These are the parameters one must specify when using Sawtooth Software CBC/HB, or the R “bayesm” “rhierMnlRwMixture” function, or almost any other software.

Looking at the R&L results for our simulated data using an initial chain for each model with 300,000 iterations, we get the following results:

- Burn-in phase length to be used ranges from 2,000 for our smallest models up to 112,000 for our larger models with heterogeneity.

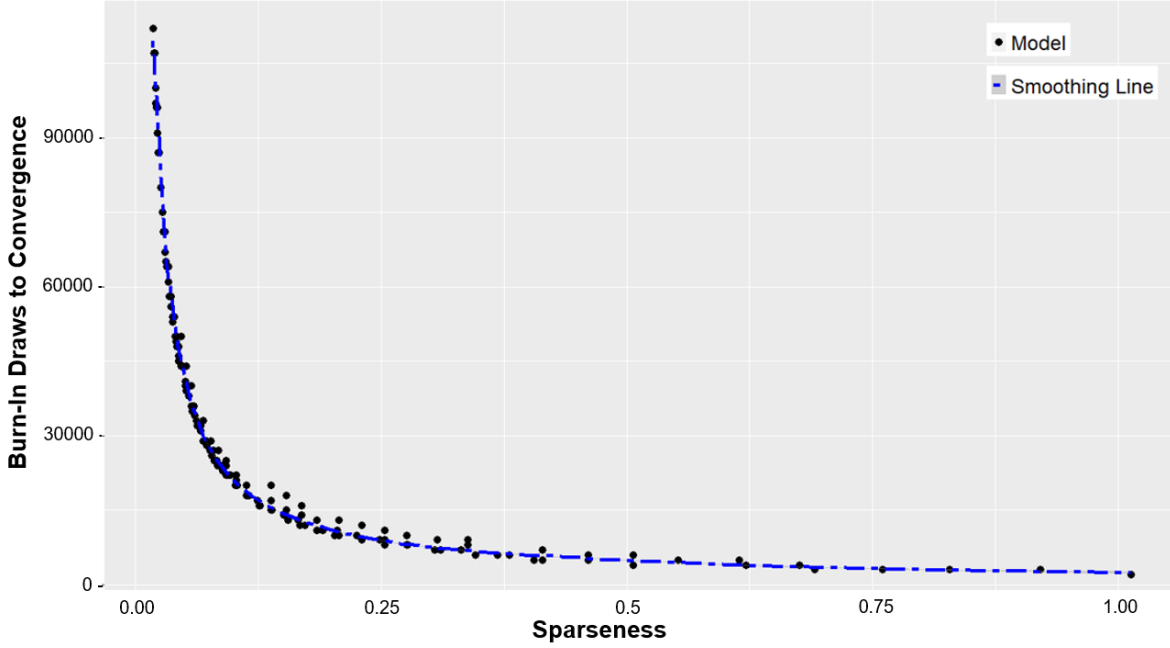
The recommended thinning interval varied from 2 to 12. We conclude that 10 is approximately right for all situations (see Figure 6).

Figure 6: Results Across All 13,824 Models



By utilizing the R&L diagnostic to determine the number of burn-in draws required for convergence, we can examine how this number behaves in relation to our sparseness index (see Figure 7).

Figure 7



Sparseness and Hit-Rate (**Smoothing line**) plotted for all 13,824 models.

The computed number of draws needed for convergence exponentially decreases as the sparseness index increases. This implies that sparse data necessitate a much higher number of burn-in iterations compared to our smaller models with denser data. Based on this relationship, it could be advantageous in practical applications to establish a rule of thumb that provides an estimate of the burn-in phase's duration until convergence occurs. Such a rule of thumb could assist in estimating the amount of time required after the completion of field work for model estimation. This is particularly valuable when dealing with complex models involving a large number of iterations.



Rule of Thumb for burn-in iterations

$$BurnIn = \frac{2000}{Sparseness}$$

$$Sparseness = nConc \times \frac{\ln(N)}{(nLevels - nAtt) \times het} \times \frac{nTasks}{100}$$

The Sparseness values estimated by this rule of thumb correlate negatively with the calculated number of burn-in draws needed based on the R&L diagnostic.



Rule of Thumb for burn-in iterations

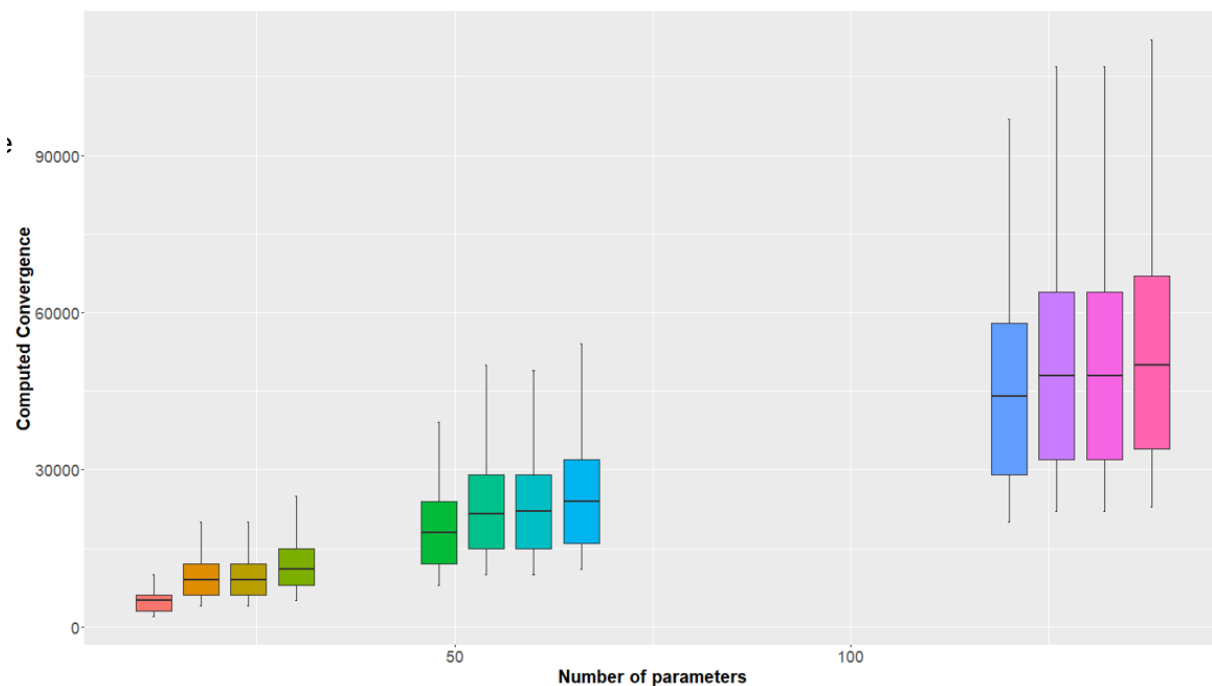
$$\text{BurnIn} = \frac{2000}{\text{Sparseness}}$$

Correlation Sparseness vs. R-L	
6 attributes	-0.611
12 attributes	-0.693
18 attributes	-0.672
24 attributes	-0.786
overall	-0.620

Correlation between the Sparseness Factor and the calculated number using the Rafferty and Lewis diagnostic.

In addition to our sparseness index, which accounts for the amount of gathered information, the number of parameters in a model is a common measure of its complexity (see Figure 8).

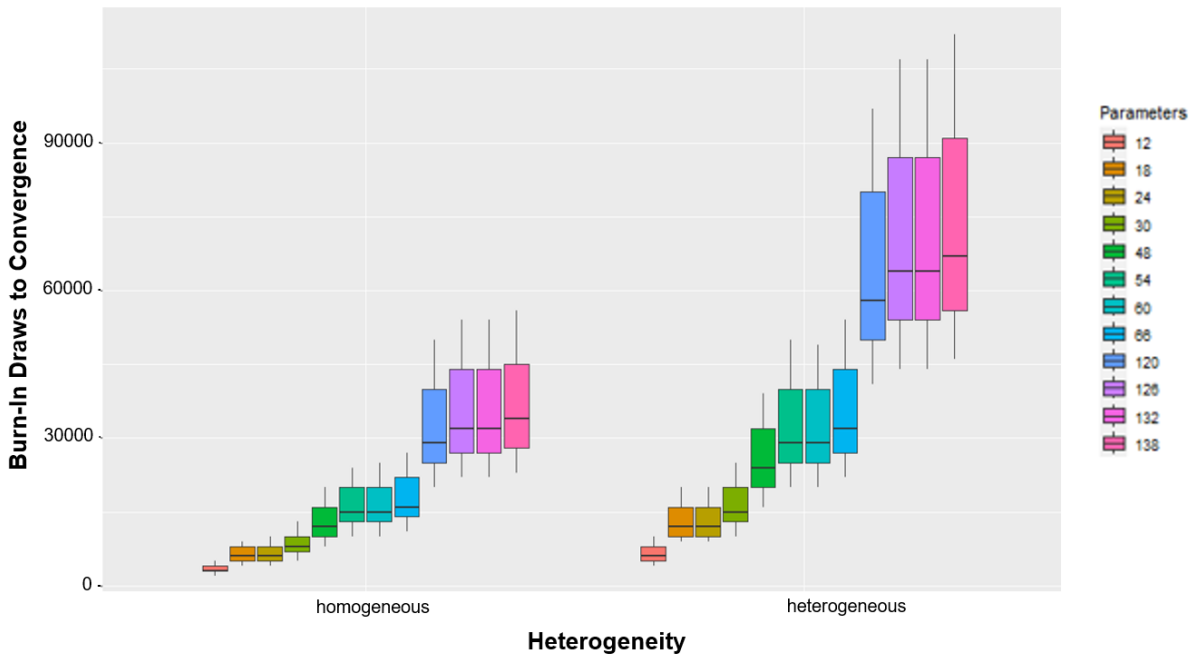
Figure 8



When plotting the number of burn-in draws required based on the R&L diagnostic against the number of parameters to be estimated, we can identify roughly three groups of models. For models with fewer than 30 parameters, 20,000 draws appear to be a suitable estimate for the burn-in phase. Models with 30 to 100 parameters require around 30,000 draws, while models with more than 100 parameters necessitate 60,000 or more burn-in draws.

Including the heterogeneity of the model in this consideration we get a more differentiated picture (Figure 9).

Figure 9



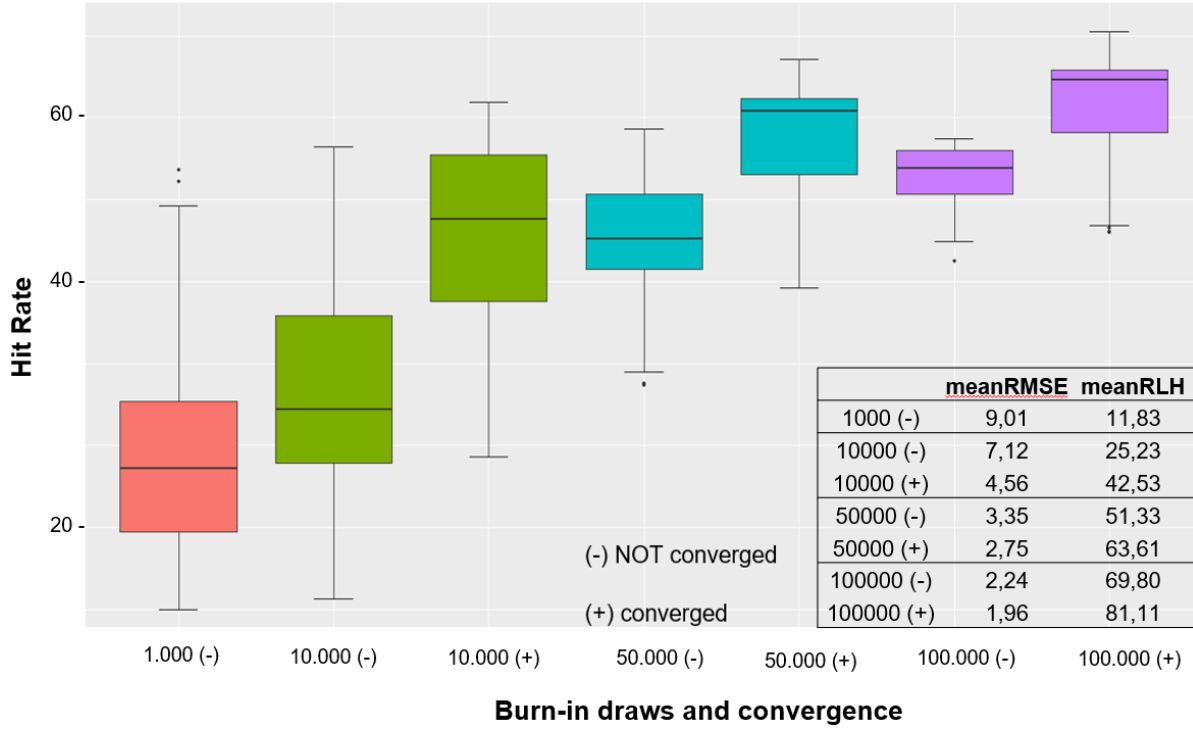
We once again observe these three groups, but we can derive different recommended values for homogeneous and more heterogeneous models (Table 2):

Table 2

	homogeneous		heterogeneous
<i><30 parameters</i>	10,000		20,000
<i>30–100 parameters</i>	25,000		50,000
<i>100+ parameters</i>	50,000		100,000

When discussing convergence and the number of iterations required to reach convergence, the question arises as to the practical implications for choice models. Analyzing the hit rates based on the number of burn-in draws and whether the respective models are considered converged or not, we find that converged models consistently exhibit significantly higher hit rates than models that did not converge (see Figure 10).

Figure 10



LONG TERM OSCILLATIONS

Finally, we examine the impact of long-term oscillations of the HB draws, particularly in the context of very sparse data. To this end, we calculated the root mean squared error (RMSE) between the true values on which the simulations were based and the simulated responses derived using the estimated part-worths.

$$RMSE(\hat{\beta}^r) = \sqrt{\frac{\sum_{n=1}^N \sum_{a=1}^A \sum_{l=1}^L (\hat{\beta}_{nal}^r - \beta_{nal})^2}{NAL}},$$

If long term oscillations appear one should see larger differences in the RMSE values between different numbers of burn-in draws (Table 3).

Table 3

n=500	10 Tasks	8 Concepts	10,000/10	10,000/100	50,000/10	50,000/100	100,000/10	100,000/100
Parameter	Heterogeneity	Burn-in	1000	100	5000	500	10000	1000
12	hom	5000	1,81	1,83	1,78	1,73	1,76	1,77
	het	10000	2,06	2,11	1,93	1,82	1,83	2,01
24	hom	10000	2,43	2,49	2,39	2,45	2,37	2,36
	het	20000	2,53	2,58	2,54	2,46	2,51	2,44
48	hom	20000	2,16	2,11	1,89	1,78	1,62	1,41
	het	39000	2,26	2,31	1,86	2,04	1,74	1,67
60	hom	25000	2,86	2,74	2,83	2,82	2,98	2,78
	het	49000	3,82	3,95	3,72	3,81	3,59	3,69
120	hom	50000	1,59	1,63	1,71	1,61	1,56	1,50
	het	97000	2,24	2,16	2,23	2,04	2,32	1,80
132	hom	54000	1,89	1,76	1,90	1,73	1,80	1,75
	het	107000	2,26	2,18	2,16	2,27	2,23	2,14
138	hom	56000	2,04	1,93	1,64	1,75	1,67	1,50
	het	112000	2,08	2,16	2,83	2,24	2,13	2,02

After the MCMC chain has converged and appropriate thinning has been applied, all RMSE measures are very similar regardless of the number of draws or higher thinning factors used. The Raftery and Lewis thinning factor, which suggests using approximately 10, proves to be suitable for all our models. In order to simplify Table 3, it only presents the smallest sample sizes and 10 tasks, representing the highest sparseness.

MCMC chains always vary around the posterior mean even after convergence, possibly resulting in oscillations of the RMSE. For the degree of sparseness covered in our models, there are no issues with long-term oscillations. However, in the case of extremely sparse data, “slow mixing” may result in oscillations and it is advisable to closely examine the results using the tools explained in this paper.

While long-term oscillations are not the main focus of this paper, we did note that in our simulation study, we did not encounter any problems related to this issue. A small indication that long-term effects could play a role can be observed in Table 3, where the RMSE values are sometimes smaller with fewer iterations than with higher numbers. This suggests that if we were to analyze significantly higher numbers of saved draws, the values would oscillate, and a higher thinning factor would be necessary. However, as these differences are minimal, we believe that digging deeper into this topic within our studied number of parameters is not necessary. For those dealing with sparser data, this aspect may become more significant.

SUMMARY

This paper addresses the question of what settings to employ when conducting a hierarchical Bayes (HB) MNL estimation, specifically regarding the number of burn-in draws and subsequent draws. We introduce a sparseness index that aids in determining the amount of information captured by the model. Based on this index, we propose a rule of thumb for the number of burn-in draws required, providing an opportunity to estimate the necessary number of burn-in draws early in the study during model development.

In general, we find that sparser models require more iterations to achieve convergence. This finding is important since models exhibit significantly higher hit rates once convergence is attained. Based on our findings, we provide ballpark recommendations for HB settings:

Parameters	Burn-In	Used Draws
<30	20,000	10,000
30-100	50,000	10,000
100+	100,000+	10,000

Rough guidance for burn-in and used draws based on our findings

For more study-specific values of burn-in draws, our rule of thumb formula based on the sparseness index can be employed. After achieving convergence and considering reasonable levels of sparseness, using 10,000 iterations while selecting every 10th draw for estimation is recommended.

For those seeking a deeper understanding of their specific models or dealing with very sparse data, the Raftery and Lewis diagnostic can be used to investigate specific settings.



Peter Kurz



Maximilian Rausch

REFERENCES

- Allenby, G.M.; Rossi, P.E. (1999):** Marketing models of consumer heterogeneity. *J. Economics*. 89 (1–2), 57–78.
- Allenby, G.M.; Rossi, P.E. (2006):** Hierarchical Bayes Models, in: Grover, R.; Vriens, M. (Eds.): *The Handbook of Marketing Research: Uses, Misuses, and Future Advances*, S. 418–440, SAGE Publications Inc., Thousand Oaks.
- Andrews, R.L; Ainslie, A.; Currim, I.S. (2008):** On the recoverability of choice behaviors with random coefficients choice models in the context of limited data and unobserved effects. *Management Science* 54:83–99.
- Brooks, S.P.; Gelman, A. (1998):** General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* 7:434.
- Gelman, A.; Carlin, J.B.; Stern, H.S; Rubin, D.B. (2008):** *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, FL.
- Gelman, A.; Rubin, D.B. (1992):** Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7 (4), 457–511.

- Goeken, N; Kurz, P; Steiner, WJ (2021):** Hierarchical Bayes conjoint choice models—Model framework, Bayesian inference, model selection, and interpretation of estimation results. *Marketing ZFP* 43:49–66.
- Hein, M.; Goeken, N.; Kurz, P.; Steiner, W.J. (2022):** Using Hierarchical Bayes draws for improving shares of choice predictions in conjoint simulations. A study based on choice data. *EJOR*, vol. 297(2), pages 630–651.
- Hein, M; Kurz, P; Steiner, WJ (2019):** On the effect of HB covariance matrix prior settings: A simulation study. *Journal of Choice Modelling* 31:51–72.
- Hein, M; Kurz, P; Steiner, WJ (2020):** Analyzing the capabilities of the HB logit model for choice-based conjoint analysis: A simulation study. *Journal of Business Economics* 90:1–36.
- Lenk, P.J.; Desarbo, W.S.; Green, P.E.; Young, M.R. (1996):** Hierarchical Bayes conjoint analysis: recovery of partworth heterogeneity from reduced experimental designs. *Market. Sci.* 15 (2), 173–191.
- Pachali, M.J; Kurz, P; Otter, T. (2020):** How to generalize from hierarchical model? *Quantitative Marketing and Economics* 18:343:380.
- Plummer, M.; Best, N.; Cowles, K.; Vines, K. (2006):** coda: Convergence Diagnosis and Output Analysis for MCMC. In *R News* (Vol. 6, Issue 1, pp. 7–11). <https://journal.r-project.org/archive/>
- Raftery, A.E.; Lewis, S.M. (1992):** One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Statistical Science*, 7, 493–497.
- Raftery, A.E.; Lewis, S.M. (1992):** How many iterations in the Gibbs sampler? In: *Bayesian Statistics 4*, eds. Bernardo, J.M.; Berger, J.O.; Dawid A.P.; Smith, A.F.M.; pp. 763–773.
- Rossi, P.E. (2022):** bayesm: Bayesian Inference for Marketing/Micro-Econometrics. <https://CRAN.R-project.org/package=bayesm>
- Wirth, R. (2010a):** Best-Worst Choice-Based Conjoint-Analysis. Eine neue Variante der wahlbasierten Conjoint-Analyse. Marburg: Tectum-Verlag
- Wirth, R. (2010b):** HB-CBC, HB-Best-Worst-CBC or no HB at all? Proceedings of the 2010 Sawtooth Software Conference 321–355