PROCEEDINGS OF THE SAWTOOTH RESEARCH CONFERENCE

October 2025

FOREWORD

These proceedings are a written report of the 26th Sawtooth Research Conference, held in New Orleans, Louisiana, May 7-9, 2025. 160 attendees participated. When Sawtooth's founder, Rich Johnson, launched the conference in 1987, he named it the Sawtooth Software Conference. In 2023 and 2024, we rebranded it as the Analytics & Insights Summit. For this 2025 conference, we changed the name to echo the original name, calling it the Sawtooth Research Conference. This new name reflects that the conference is not a software training event, but a research conference encompassing diverse quantitative methods and software tools.

We appreciate the financial support of our 2025 conference sponsor, **Knowledge Excel**. Their generosity helped offset the costs of hosting a high-quality in-person event. Sponsorships did not influence paper selection for the main sessions or these proceedings, which followed an open call and rigorous review process, thereby maintaining the educational integrity of the program.

The Sawtooth Research Conference remains focused on quantitative methods in marketing research. The authors were charged with delivering presentations of value to both the most sophisticated and least sophisticated attendees. Topics included pricing research, the use of AI in quantitative research, choice/conjoint analysis, modeling/predicting market share, MaxDiff, market segmentation, and classification.

The papers and discussant comments are in the words of the authors and very little copyediting was performed. At the end of each paper are photographs of the authors and co-authors. We appreciate their cooperation in providing photos, which lend a personal touch and help readers recognize them at future conferences.

We are grateful to these authors for continuing to make this conference such a valuable event. We believe the Sawtooth Research Conference 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 especially grateful for the efforts of our steering committee, whose continued dedication makes this conference a success: Keith Chrzan, Marco Hoogerbrugge, Joel Huber, David Lyon, Ewa Nowakowska, Bryan Orme (Chair), and Megan Peitz.

Sawtooth Software

October, 2025

CONTENTS

SUMMARY OF FINDINGS

The 26th Sawtooth Research Conference was held in New Orleans, Louisiana, May 7–9, 2025. The summaries below capture some of the main points of the presentations and provide a quick overview of the articles available within the 2025 Sawtooth Research Conference Proceedings.

Blending Historical MaxDiff Claim Studies and Using AI to Predict Claim Success (Jeremy Christman, P&G, Kevin Lattery, Nino Hardt, SKIM, Liz Clevenger, David Hengehold, Pankaj Patil, P&G and Howard Huang, SKIM): Jeremy and his co-authors introduced an AI-driven framework that unified 73 independent MaxDiff studies at P&G—spanning 1,554 product claims and 44,000 respondents—to predict how new consumer packaged goods (CPG) claims would perform. A "glue study" anchored all results onto a common scale, enabling integration through a hierarchical Bayesian model with a factored covariance structure that preserved 96% of variance while cutting computation time by two-thirds.

Using OpenAI's 3,072-dimensional text embeddings and AutoGluon AutoML, the model predicted each claim's "beat-average probability" (BAP), or likelihood of outperforming an average claim. The embedding-based AutoML model achieved high accuracy (R^2 = 81.4%), outperforming GPT-style prompting and simpler algorithms. This approach transformed fragmented testing data into a unified, predictive system that allowed P&G to evaluate new claim ideas, reduce redundant studies, and accelerate innovation. While the model's accuracy depends on existing semantic coverage and does not yet account for time-based shifts in preferences, it provides a scalable foundation for AI-enhanced claim analysis and forecasting in marketing research.

AI At Microsoft: Enhancing Conjoint and Scaling In-Depth Interviews (Daniel Penney, Suhasini Sanyal Saxton, Ananya Ramje, Microsoft, Evan DiMarzio, PSB Insights): Microsoft and PSB Insights studied how AI can enhance market research by being integrated into conjoint (as AI-driven probes following open-ended comments) and for scaling in-depth interviews (IDIs). In experiment 1, in which AI probes were added to a conjoint survey, response length to open-ended questions increased by 57%, vague answers were reduced from 45% to 6%, and model fit improved from 90% to 96%. It also showed potential to better predict real-world behaviors. Experiment 2 compared text and voice responses to open-ends. Voice increased the length of open-ended response by 21% but tended to include more conversational filler, reducing their overall utility for modeling purpose; model fit remained 89%. The authors suggested that voice response suited exploratory studies, whereas text responses better suited structured surveys. Experiment 3 tested AI-moderated interviews. AI replicated many benefits of human moderators, reduced time and cost, and provided a non-judgmental, asynchronous format, though it struggled to probe contradictions or read emotional cues. Overall, AI can improve engagement, data richness, and data quality in surveys, and in some cases, can be a helpful augment or alternative to traditional qualitative research – but its use needs to be assessed on a case by case basis to determine its appropriateness.

Like to Get to Know You Well: Strategies for Streamlining Variables to Create Actionable and Engaging Segmentation Models (Tracey Di Lascio-Martinuk, Bose Corporation): Tracey's team at Bose developed a consumer segmentation model for the premium audio market by focusing on early strategic decisions and stakeholder engagement. They first identified key stakeholders, clarified use cases, and aligned expectations, which guided variable selection and model design. Initial input included a broad set of demographic, behavioral, and psychographic variables, but the team streamlined dimensions to avoid multicollinearity, overly narrow segments, and poor differentiation. Survey questions were carefully designed to capture nuanced traits, and variables were reviewed for correlation, response spread, and meaning before modeling. The team used latent class analysis to develop segments that balanced statistical rigor, explainability, and business actionability. Once defined, segments were applied through a typing tool, commercial database mapping, and media targeting. By limiting dimensions and focusing on practical applications, the team produced a model that stakeholders trusted, could locate in real-world data, and used effectively across business decisions, highlighting the importance of strategic planning alongside technical modeling in segmentation.

(Sponsorship did not influence paper selection for the main sessions or these proceedings)

Evaluating Mobile-Friendly Conjoint Designs (Paul Richard McCullough, MACRO Consulting, Inc. and Dan Yardley, Sawtooth Software): Paul and Dan compared three conjoint methods—Traditional CBC (Carousel format), Tinder CBC (single-alternative swipe), and Slider Conjoint (single-alternative slider scale)—across desktop and mobile environments, focusing on model performance, respondent engagement, and dropout rates. For this comparison, they surveyed 1,500 U.S. respondents using balanced experimental designs and consistent holdouts. Traditional CBC (with carousel format for mobile view) outperformed both single-alternative formats, achieving the highest hit rate, lowest out-of-sample MAE, lowest dropout rate, and lowest attribute non-attendance, indicating deeper respondent engagement and more robust model estimates. Tinder and Slider formats yielded slightly higher enjoyment and shorter completion times but showed lower model robustness and higher rates of flatlining responses. Structural equation modeling revealed that survey clarity ("easy to understand") significantly influenced model performance, whereas enjoyment did not. Paul and Dan concluded that Carousel CBC should be preferred for high-stakes, accuracy-critical applications, while Tinder CBC offers a valid alternative when respondent experience is prioritized. Findings highlight the trade-offs between data quality and user engagement in mobile-first survey design.

*** Token-Based Conjoint—A New Framework for Too Many Attributes (Megan Peitz and Trevor Olsen, Numerious, Inc.):** Token-Based Conjoint (TBC) addresses studies with large feature sets (30–40 attributes) where traditional conjoint and MaxDiff struggles. It combines MaxDiff and choice-based conjoint, asking respondents to allocate tokens to top features in small subsets, followed by a purchase likelihood question. This captures relative appeal and purchase thresholds while enabling additive modeling.

Megan and Trevor discussed token-based conjoint studies they've conducted, where respondents viewed n features per screen, selecting k favorites, with k varying across tasks to reflect real-world trade-offs. Responses were treated as discrete choice tasks for hierarchical Bayes or multinomial logit modeling. Validated across three studies, TBC showed strong predictive accuracy. Naïve none coding improved purchase predictions, and weighted likelihoods corrected overrepresentation in augmented tasks. The authors reported that TBC scales well, engages respondents, and produces actionable outputs like feature rankings and uptake simulations. It excels in high-dimensional, additive-value studies, effectively bridging the gap between traditional conjoint and MaxDiff methods, but is less suited for price-focused research.

* Co-recipient of Best Paper Award as voted by the audience

Synthetic AI Avatars in Market Research–A Game Changer or a Mere Gimmick? (Saurabh Aggarwal, Tarun Khanna, and Rashmi Sharma, Knowledge Excel): Saurabh and co-authors evaluated synthetic AI avatars in online surveys for their ability to improve engagement, data quality, and respondent experience. Traditional text surveys face declining attention and reliability, so avatars—video agents that explain, express and connect—were tested in three phases: feasibility, prototype refinement with 187 research professionals, and a field experiment with 1,668 respondents in India. Avatars were deployed at varying levels: instructional-only, avatar selection with toggle-control, and continuous presence.

Moderate avatar use increased engagement, emotional and cognitive involvement, and thoughtful open-ended responses. Avatars reduced low-effort behaviors like straightlining and speeding and improved MaxDiff task consistency without harming predictive accuracy. Respondents preferred female avatars, which elicited higher engagement and comfort. Providing toggles and persona choice preserved autonomy and reduced fatigue, while continuous exposure slightly lowered completion rates.

Feedback confirmed avatars clarified instructions and humanized surveys, without triggering uncanny-valley effects. Practical deployment guidelines emphasized transparency, cultural adaptation, technical performance checks, and user control. Looking forward, intelligent LLM-powered avatars could enable two-way adaptive surveys, integrating qualitative and quantitative data collection. Overall, AI avatars show promise to enhance survey engagement and data quality when applied thoughtfully.

Transforming Open-Ended Survey Response Analysis with AI Transformer Models (Jacob Nelson, The Harris Poll): Open-ended survey questions capture rich, nuanced insights but are slow and costly to analyze manually. Jocob described how traditional text analysis methods like Term Frequency–Inverse Document Frequency (TF-IDF), which scores word importance based on frequency, and Latent Dirichlet Allocation (LDA), which uncovers latent themes from word co-occurrence, can struggle with context and semantic nuance in short survey responses. Transformer-based AI models—embeddings and large language models (LLMs)—offer scalable solutions. Embeddings convert text into vectors to cluster semantically similar responses, while LLMs classify and abstract responses into broader themes.

Jacob reported on a study conducted by The Harris Poll involving 10,000+ U.S. vehicle owners, key features were extracted from responses about desired automotive technologies. Embeddings produced 224 fine-grained codes, and LLMs generated 129 broader, conceptual codes. Both achieved 75–81% agreement with human coding. Embeddings excel at grouping similar ideas expressed differently, while LLMs are better at forming hierarchical, interpretive themes. Jacob concluded, optimistically, that AI-assisted workflows reduce manual effort and make open-end analysis scalable and practical.

Deepening Consumer Insights—Conversational AI (Mohit Shant, Rajat Narang, Md. Faisal, Insights Curry): The authors conveyed how conversational AI is reshaping market research by replacing static, one-directional surveys with dynamic, adaptive dialogues that enhance engagement, trust, and decision-making. Their research examined Agent Mira, a multi-agent conversational AI embedded in a real estate app, tested with 154 active homebuyers. Mira guided users through property evaluations, clarified questions, and provided personalized insights such as offer prices and win probabilities. Results showed 99% engagement, 81% influence on decisions, and 83% of initially skeptical participants willing to use AI again.

Mohit and team asserted that by enabling context-aware, iterative interactions, conversational AI addresses limitations of traditional surveys, including low engagement, ambiguous questions, and inattentive or bot responses. Key design features included specialized agents for domain expertise, ethical guardrails, contextual memory, and continuous learning loops. Follow-up prompts and adaptive questioning further improved user satisfaction and trust. These methods improved data richness, reduced low-quality responses, and fostered more humanized interactions. Beyond real estate, applications extend to healthcare, finance, product development, and public services, offering a scalable approach to gather deeper, higher-quality consumer insights and influence strategic decision-making.

Revisiting the No-Choice Option in Conjoint Analysis (Cheng-Yu Hung, Ohio State University, Peter Kurz, bms marketing research + strategy, Roger A. Bailey, Ohio State University, Joel Huber, Duke University, and Greg M. Allenby, Ohio State University): Cheng-Yu and co-authors reviewed the role of the no-choice option in conjoint analysis, emphasizing its importance in capturing realistic consumer preferences and economic valuation. The no-choice option allows respondents to opt out of purchasing any of the presented alternatives, reflecting the utility of outside options. The authors show that when respondents lack prior knowledge of market prices or offerings, as in an MP3 player study conducted many years ago, they dynamically update their preference for the outside option throughout the choice tasks. Brand part-worths are affected by this learning, while other product attributes remain stable. Conversely, in mature categories such as tooth-whitening products, where respondents are familiar with prices and product features, the authors found that dynamic updating is minimal, and providing market information as training prompts further stabilizes outside good preferences.

The authors reported on two studies that illustrated these effects: a new-product MP3 player study across France, Italy, and the UK, and a mature-product tooth-whitening study in the U.S. Models incorporating dynamic updating of the outside good better fit the data than standard models. The findings highlight that including the no-choice option is essential for accurate demand estimation and that properly screening respondents, clearly defining product features, and optionally providing market context mitigates biases from learning. Overall, brand values are sensitive to familiarity and learning, but other attribute-level utilities are robust, supporting effective conjoint study design.

Optimizing Product Portfolios with Discrete Choice Models: A Practical Approach for Market Researchers (Maximilian Rausch, Peter Kurz, and Stefan Binner, bms marketing research + strategy): Maximilian and his co-authors presented a framework for optimizing FMCG product portfolios using discrete choice models. Portfolio optimization is complex due to limited shelf space, diverse consumer preferences, and multiple strategic goals, such as maximizing revenue, market share, or profit. The paper outlines simple stepwise methods for small SKU sets and systematic approaches for larger sets, incorporating business constraints like must-have SKUs, production costs, and portfolio size limits.

The authors discussed three optimization techniques: exhaustive search, simulated annealing (SA), and genetic algorithms (GA). Exhaustive search guarantees the global optimum but is computationally heavy. SA explores solutions probabilistically to avoid local optima, while GA evolves high-performing portfolios efficiently. Hybrid methods combining GA with focused exhaustive search can reduce computation time while maintaining quality. The authors emphasized that mathematical optima may conflict with practical constraints. A collaborative optimization framework integrates quantitative results with managerial judgment, using top-performing portfolio profiles and business logic to design realistic, strategically sound assortments, balancing analytical rigor, efficiency, and market feasibility.

Toward a Smarter MaxDiff: Rethinking Some Conventional Strategies (Ming Shan, Hall & Partners): Ming explored an adaptive MaxDiff approach that adapts survey design in real time using insights from Item Response Theory (IRT) and Computerized Adaptive Testing (CAT). Traditional MaxDiff surveys are sequential—design, data collection, then modeling—often using balanced designs and hierarchical Bayes (HB) estimation. However, balanced designs can underperform, particularly for mid-tier items, as many tasks become too easy and provide limited information.

Shan proposed modeling respondents individually during the survey, updating item exposure dynamically based on early responses and previously completed tasks. IRT is used for fast, interim estimation of item importance, while CAT concepts guide adaptive task selection, improving precision and efficiency. Simulations show that overexposing high- or mid-importance items enhances accuracy, reduces required sample sizes, and improves respondent-level precision compared to standard MaxDiff. Ming's research highlights the potential of unbalanced, data- and goal-dependent designs. It demonstrates that adaptive, across- and within-respondent learning can substantially improve MaxDiff performance. Future work could refine multi-dimensional, respondent-level adaptation and optimization algorithms for broader applicability.

Using Conjoint to Assess the Value of Brand Associations (Marco Vriens and Felix Eggers, Kwantumlabs.ai): Marco Vriens and Felix Eggers explored using choice-based conjoint analysis to assess the value of brand associations, comparing traditional pre-defined perception ratings with open-ended responses. Traditionally, brands are included as attributes in conjoint studies to estimate utility and derive dollar-equivalent brand equity. Alternatively, brand perceptions can be modeled as predictors of brand utility, though these ratings may suffer from low correlation with market share, halo effects, or response biases, and often overlook negative associations.

The authors propose using open-ended questions, such as "What comes to mind when you think of brand X," to capture authentic brand associations. Responses are coded into distinct associations, which can now be analyzed with AI/text analytics to quantify their frequency and strength. These associations are then integrated into conjoint models as binary predictors to estimate their contribution to brand utility and monetary value. Empirical studies show that open-ended associations strongly influence consumer surplus and brand choice, often more intuitively than pre-defined perceptions. Positive associations increase willingness-to-pay, negative associations reduce it, and the total number of associations—reflecting mental availability—also enhances brand value. Open-ended responses thus offer richer, actionable insights into brand equity.

AI-Assisted Segmentation: Better, Faster, and More Actionable Segmentation (Jackie Guthart, Curtis Frazier, Marcos Nuñez, Yamil Bongioanni, and Marrissa Simons, Radius Global Market Research): Jackie and co-authors presented an AI-assisted hybrid workflow for market segmentation at Radius Global Market Research, aiming to make segmentation faster, more actionable, and interpretable. Traditional segmentation is time-intensive, requiring analysts to manually evaluate numerous solutions, craft segment profiles, and compare options—tasks prone to cognitive overload and inefficiency. The hybrid approach combines human expertise with AI capabilities across three components: a Segmentation Development Assistant, AI Segment Evaluation and Naming Tool, and AI Segment Comparison Assistant.

The Segmentation Development Assistant uses a scoring algorithm incorporating segment distribution, variance, and differentiation to prioritize high-quality solutions. The AI Segment Evaluation and Naming Tool automates generation of segment names, descriptions, and key attributes, freeing analysts to focus on strategic interpretation while maintaining 94.5% overlap with human insights. The AI Segment Comparison Assistant identifies similarities, differences, activation readiness, and segment value, even filling gaps in survey data. The authors reported an over tenfold efficiency improvement, enriched insights, and more actionable recommendations. They conclude that a hybrid AI-human model delivers faster, higher-quality segmentation while maintaining oversight and interpretability, with ongoing enhancements expected for AI context integration and client-facing interactivity.

Segmentation 2.0: Redefining Segmentation for Modern Marketing (Catherine Gibson, Jessica Wojtunik, Alexandra Chirilov, Agnieszka Fronczyk, Rachel Thompson, and Irina Nazarova, NIQ): Catherine and co-authors proposed a modernized framework for market segmentation, addressing the tension between targeted marketing and mass-market brand building. Traditional Segmentation, Targeting, and Positioning (STP) models remain useful for identifying consumer differences and enabling short-term activation, but contemporary marketing theory emphasizes broader market penetration, mental and physical availability, and long-term brand growth, as highlighted by Byron Sharp and the Ehrenberg-Bass Institute.

The authors introduced "Segmentation 2.0," which extends conventional segmentation through three analytical enhancements: (1) Upscaling—leveraging pairwise similarities to expand insights from core target segments to adjacent audiences; (2) Broad Reach Targeting—aggregating segments into thematic platforms to enable scalable communication; and (3) Brand Building—identifying universally resonant themes that support mass-market messaging. This dual approach balances short-term activation with long-term brand equity, ensuring insights from differentiation coexist with insights from commonalities. Methodologically, the framework uses similarity matrices, hierarchical cluster analysis, and decision rules inspired by game theory (e.g., MaxiMin) to optimize both segment-specific and market-wide strategies. Segmentation evolves from a static tool into a dynamic process that informs scalable, flexible marketing strategies, bridging the gap between niche targeting and broad-reaching brand growth.

Modernizing Data Visualization Practices for Market Research (J. Keaton Wilson and Ben Cortese, KS&R): Modern market research generates vast amounts of data, yet visualization practices often rely on outdated formats like bar and pie charts, which can obscure insights. Keaton and Ben discussed how effective visualizations must prioritize clarity, accuracy, and accessibility, minimizing cognitive load and enabling data to "speak" on its own. Drawing on research from graphical perception, cognitive science, and design, modern visualization emphasizes perceptually accurate encodings—favoring position and length over angles or color—and avoids distracting elements or "chartjunk." Cognitive principles highlight working memory limits, recommending simplified, contextually guided visuals, while accessibility considerations ensure colorblind-safe palettes and redundant cues improve comprehension for all viewers. Minimalist, honest design—following Tufte and Kosslyn—maximizes data communication without distortion.

The authors showed practical examples to demonstrate these principles: proportional-width histograms and treemaps correct misinterpreted bins, violin plots reveal distributions hidden by bar averages, lollipop charts outperform pie charts for part-to-whole comparisons, and radar charts synthesize multidimensional segment profiles. Adoption requires balancing best practices with tool constraints, stakeholder expectations, and data literacy through incremental implementation, training, and clear communication. Modernized visualization enhances insight delivery, supporting better decisions and more effective storytelling in market research.

Turning it to 11: A Practitioner-Led Comparison of Volumetric Conjoint Analysis Techniques (Dean Tindall, Sawtooth Software, Chris Moore, and Manjula Bhudiya, Ipsos): Volumetric conjoint analysis extends traditional Choice-Based Conjoint (CBC) by not only identifying preferences but also estimating the quantities respondents are likely to purchase, providing insights into market demand, pricing, and operational planning. Despite its value, Dean and co-authors commented that it remains underutilized due to increased survey complexity, data variability, and modeling challenges. Their study evaluated five volumetric modeling methods—Naïve, MEV (Maximum Expected Value), Joint Discrete/Continuous (JDC), Polytomous Logit (MBC), and HB-Regression—using a mid-complexity UK potato crisps study with 2,003 respondents. The authors emphasized that volumetric tasks are cognitively demanding, prone to over- or under-reporting, positional bias, and device effects, requiring rigorous data cleaning, capping of extreme values, and attention to early-task anomalies.

The authors found that simpler methods like Naïve performed well for stable, moderate-sized SKU scenarios, while JDC and MBC offered stronger, scalable performance across larger SKU sets (extrapolating from the calibration sets used for building the models). HB-Reg showed instability under complex scenarios. Exponent (scale factor) tuning further improved accuracy, except for the HB-Reg model. Key recommendations include prioritizing data quality, accounting for device effects, discarding early tasks, selecting models aligned with study complexity, and calibrating utilities. The authors concluded that the value of volumetric conjoint lies in balancing methodological sophistication with practical feasibility, with good data quality being critical to success.

*** The Will of the Many: Generating Novel Concepts Using AI-Enhanced Respondent Feedback (Joris van Gool and Peter Li, SKIM):** Joris and Peter presented a novel methodology for new product design (NPD) that leverages AI and respondent input to generate innovative, aesthetically-appealing concepts. Traditional survey-based methods like conjoint analysis often struggle in creative contexts due to the vast and high-dimensional design space, which can exclude attractive consumer ideas. To overcome this, the authors allowed respondents to describe their ideal designs in natural language, which were then transformed into images using state-of-the-art text-to-image AI models. These AI-generated designs were evaluated alongside professional designs (from human designers) through a large-scale MaxDiff survey, using Christmas-themed Coca-Cola cans as a case study.

Their methodology combined exploration of the design space with exploitation of high-potential designs, while large language models (LLMs) automated prompt rewriting and attribute extraction from both text and images. Results show that respondent-driven AI designs cover a wide variety of creative ideas and often outperform professional designs at the upper end of preference. Joris and Peter's study demonstrated scalable, automated modeling of consumer preferences in a highly complex and visual design space. Future research could improve image fidelity, prompt adherence, and feature identification for even more effective NPD.

* Co-recipient of Best Paper Award as voted by the audience

Better Segmentation Results with Deep Learning: Dimensionality Reduction Using Auto-Encoders (Joseph Retzer, ACT-Solutions): Joseph compared Principal Component Analysis (PCA) and deep learning–based Auto-encoders (AEs) for dimensionality reduction in clustering tasks. Using a dataset of 28 numeric features, both methods reduced the data to seven dimensions before applying the Partitioning Around Medoids (PAM) clustering algorithm. Cluster quality was evaluated with the Calinski-Harabasz Index (CHI), Davies-Bouldin Index (DBI), and Silhouette scores. Results showed that AEs significantly outperformed PCA and raw data, achieving a silhouette score of 0.32—well above the interpretability threshold.

Auto-encoders captured nonlinear relationships and preserved more information, producing clearer, more meaningful clusters. Visualization tools such as correlation and dependence heatmaps and feature permutation importance plots enhanced interpretability of AE components. While AEs required more tuning (e.g., optimizing learning rates, layers, and activation functions), their flexibility and superior performance justified the added complexity. Joseph's work also highlighted R's tidymodels framework as a reproducible tool for feature engineering tasks like imputation, normalization, and transformation. Overall, deep learning–based AEs offered substantial advantages over PCA for dimensionality reduction, yielding higher-quality segmentation and more insightful unsupervised learning outcomes.

Enhancing Cluster Ensembles with Latent Class Clustering (Keith Chrzan, Sawtooth, and Joseph White, Kynetec): Keith and Joseph investigated whether adding latent class clustering—a model-based approach assuming data arise from mixtures of probability distributions—can improve clustering performance in Sawtooth Software's CCEA (Convergent Cluster and Ensemble Analysis) for a variety of synthetic datasets. CCEA aggregates multiple clustering solutions (70 by default, across 14 segment sizes and several algorithms) into a consensus segmentation. The researchers enhanced CCEA by incorporating mclust and Latent Gold latent class solutions into the custom ensemble and compared results using 47 synthetic data sets with known segment structures.

They evaluated performance using two criteria: (1) accuracy in recovering the true number of clusters and (2) respondent classification accuracy measured by the Adjusted Rand Index (ARI). Results showed that while the enhanced ensemble (CCEA + mclust) improved upon default CCEA, mclust alone typically outperformed both. Across studies, mclust identified the correct number of clusters more often and achieved higher ARI values. Mclust performed slightly better than Latent Gold for these synthetic datasets. Overall, the research concludes that model-based clustering methods like mclust or Latent Gold provide superior segmentation performance and should be considered primary tools, though enhancing ensemble methods for CCEA can still yield modest improvements over its default ensembles.

Determining the Value of Price Thresholds in Pricing Conjoint Studies (Michael Smith and Juli Pham, SKIM): Michael and Juli examined whether modeling price thresholds—points where consumer demand drops sharply—improves conjoint pricing models. They compared two methods across 8–12 studies: Post-Hoc Cliffing, which inserts strongly negative utilities above each respondent's maximum observed chosen price, and Cliffing in Estimation, which estimated the penalty at and beyond that maximum observed chosen price in the HB-MNL model. Studies covered categories like consumer goods, software, and personal care, using both slope and discrete price codings.

Results showed that neither method meaningfully improved model accuracy compared to the default part-worth (effects-coded) approach for modeling price. Across metrics such as hit rate, mean absolute error, log-likelihood, and correlation with market data, differences were negligible—typically below 1–2%. However, Post-Hoc Cliffing slightly outperformed the modeled version, showing minor gains in predictive accuracy and modest reductions in willingness-to-pay (WTP). It also slightly increased the relative impact of price in simulations, helping temper inflated WTP estimates. Practically, clients appreciated Post-Hoc Cliffing for creating more intuitive "cliffs" in demand curves, though its success depended on testing a wide enough price range. Overall, price thresholds offered interpretive value but limited quantitative improvement—best used as a diagnostic tool rather than a required model enhancement.

Leveraging the 4 P Marketing Framework to Calibrate Conjoint Models (James Pitcher Dimitri Liakhovitski, and Alexandra Chirilov, Nielsen NIQ): James and co-authors reported on whether incorporating all elements of the 4P marketing framework—Product, Price, Place, and Promotion—could improve the calibration of conjoint models, aligning them more closely with actual market shares. Traditional calibration typically adjusts for Place (distribution), since by default conjoint analysis assumes full availability. Using NielsenIQ Point-of-Sale data and conjoint surveys across two product categories (chocolate bars and large-screen TVs) in five countries, James and his team tested multiple calibration methods, including scale factor adjustment (to bring respondents' response error more in line with real-world buyer behavior), direct 4P adjustments, and regression-based approaches.

Results showed that calibration using distribution (place) and scale factor significantly reduced mean absolute error (MAE) between conjoint and market shares. Adding the other three Ps occasionally improved accuracy—especially when Price or Product metrics were included—but gains were inconsistent across markets. Regression models confirmed that the non-Place Ps contained useful predictive information, though post-hoc adjustments risked distorting market simulation relationships among the SKUs in terms of own-price and cross-price elasticity. Survey-based 4P metrics correlated well with sales data for Product and Place but less for Price. Overall, incorporating the full 4P framework showed theoretical and practical promise for improving conjoint calibration, though further research is needed to refine weighting and respondent-level integration.

BLENDING HISTORICAL MAXDIFF CLAIM STUDIES AND USING AI TO PREDICT CLAIM SUCCESS

JEREMY CHRISTMAN
P&G
KEVIN LATTERY
NINO HARDT
SKIM
LIZ CLEVENGER
DAVID HENGEHOLD
PANKAJ PATIL
P&G
HOWARD HUANG
SKIM

ABSTRACT

In consumer packaged goods (CPG) industries, product claims are crucial for communicating benefits to consumers. This work presents a novel approach to integrate multiple historical MaxDiff studies and leverage artificial intelligence to predict claim success. We address three key methodological challenges: (1) combining independent MaxDiff studies into a unified framework, (2) fitting large-scale choice models using factored covariance matrices, and (3) developing machine learning models to predict performance of new, untested claims. We demonstrate this approach through an empirical application involving 73 MaxDiff studies comprising 1,554 unique claims tested across approximately 44,000 respondents. Our methodology involves conducting a "glue" study with strategically selected claims, implementing a hierarchical Bayesian model with factored covariance in Stan, and training embedding-based neural networks for claim prediction. Results from our application demonstrate that the embedding-based AutoML approach achieves 81.4% R-squared in predicting beat average probability for new claims, enabling instantaneous evaluation of potential product claims. This integrated approach transforms serial, independent claim testing into a comprehensive claims database with predictive capabilities, significantly accelerating the product development cycle.

1. INTRODUCTION

Product claims play a vital role in consumer packaged goods (CPG) marketing, helping communicate product benefits and differentiate offerings in competitive markets. For categories such as paper towels, laundry detergents, and oral care products, the space of potential claims is vast, encompassing various benefit statements, semantic variations, and stylistic approaches. Examples include "2X More Absorbent* So You Can USE LESS" for paper towels or "50% MORE fresh*" for laundry detergents (see Figure 1).

Continuous experimentation and testing are essential for identifying effective claims that resonate with consumers. At Procter & Gamble (P&G), the widespread availability of research through DIY tools has enabled product teams to conduct MaxDiff studies frequently and independently. While this has generated a wealth of data—with studies running constantly across multiple product categories throughout the year—it has also created challenges.

Studies are executed independently by different teams to meet their specific objectives and timelines, which results in siloed insights that cannot be easily shared or leveraged across the organization. Each study produces relative preference scores that cannot be directly compared across studies, making it impossible to establish which claims perform best in absolute terms. There is no systematic way to leverage learnings from past studies when designing new ones, leading to potential duplication of effort and missed opportunities to build on previous insights. Furthermore, teams cannot quickly assess whether a new claim idea has already been tested or how it might perform relative to previously tested alternatives, creating inefficiencies in the research process.

This paper addresses two interconnected objectives: (1) integrating multiple historical MaxDiff studies into a unified claims database, and (2) leveraging artificial intelligence to enable instantaneous prediction of new claim performance. Our approach transforms isolated preference studies into an integrated prediction system, significantly accelerating the product development cycle.

Figure 1

2. Challenges and Considerations

To achieve the two goals of consolidating past studies and creating an instantaneous prediction engine, we need to consider a few challenges of meta-analyses, predictive modelling and incorporating language models.

2.1 Meta-Analysis

Integrating many independent (un-anchored) MaxDiff studies into a unified framework requires addressing fundamental differences in how these studies were designed, executed, and analyzed. While each study provides valuable insights about consumer preferences within its specific context, combining them to create a comprehensive claims database introduces complexities that go beyond traditional meta-analytic approaches. In return, leveraging a large database of claims covering a variety of benefits and linguistic devices helps inform predictive models.

A key challenge is that MaxDiff studies produce relative preference scores within each study. Without common anchoring across studies, scores from different studies cannot be directly compared. A claim scoring 80 in one study may represent different absolute preference than a claim scoring 80 in another study. This fundamental limitation prevents us from establishing which claims perform best in absolute terms across the entire corpus of research.

Compounding these challenges is the minimal overlap in claims tested across studies. Among the 1,554 unique claims in our dataset, most appear in only one study, preventing direct bridging through common items. This lack of overlap means traditional meta-analytic approaches that rely on shared items across studies cannot be applied.

Finally, unlike traditional conjoint analysis where products can be decomposed into attributes and levels, claims lack a natural factorial structure. Each claim is essentially unique, making it difficult to define features or factors for modeling. While it might be possible to “engineer” such a structure by decomposing claims into linguistic features or benefit categories, this approach brings its own challenges and may not capture the nuanced ways that specific wording drives consumer preference. Given our prediction focus, we pursue an alternative path that leverages modern NLP techniques to capture semantic meaning directly from claim text, which proves more effective for our predictive objectives.

2.2 Predictive Modeling

Beyond integrating historical data, predicting performance for new claims presents a distinct set of challenges that stem from the fundamental nature of marketing language and consumer psychology. Claims are free-form text that vary dramatically in length, style, and content—from concise benefit statements like “2X More Absorbent” to elaborate descriptions incorporating multiple benefits and emotional appeals. This textual complexity makes traditional feature engineering approaches inadequate for capturing the nuanced differences that drive consumer preferences.

The relationship between claim language and consumer preference exhibits extreme non-linearity. Small wording changes can have disproportionate impacts on performance: “2X More Absorbent” and “Twice the Absorption Power” convey similar functional benefits but may perform very differently due to subtle differences in perceived credibility, memorability, or emotional resonance. Conversely, claims that appear quite different on the surface may perform similarly if they tap into the same underlying consumer need or motivation. This complexity requires sophisticated modeling approaches that can capture both semantic meaning and the psychological impact of specific word choices.

Furthermore, predictive models must generalize beyond the patterns present in historical data. New claims often introduce novel concepts, benefits, or linguistic patterns that weren't present when the model was trained. For instance, sustainability claims have evolved rapidly in recent years, with new terminology and benefit framings emerging constantly. The model must be able to evaluate these novel claims intelligently rather than simply memorizing patterns from historical data. This requires a representation that captures the underlying semantic and psychological dimensions of claims rather than relying on surface-level features.

2.3 Other Considerations

Instead of leveraging existing preference data, some practitioners might feel tempted to rely on "synthetic respondents" to predict performance of new claims. These approaches come under a variety of names (e.g., "synthetic samples," "silicon samples") with varying definitions and implementations. The general approach attempts to replace human respondents with LLM-simulated responses, sometimes conditional on persona descriptions.

However, synthetic respondent approaches face significant challenges. Some research calls into question the ability of LLMs to simulate human behavior (e.g., Gao et al., 2025). More specifically, though, there are domain-specific information gaps. While LLMs understand basic product categories, they lack the nuanced understanding of what makes specific claims resonate with consumers. Second, existing domain-specific data is difficult to incorporate, and calibration of synthetic responses to established KPIs is difficult. Third, current LLMs are tuned to be "helpful assistants" rather than "representative respondents," suffering from sycophancy and behaving differently from human respondents.

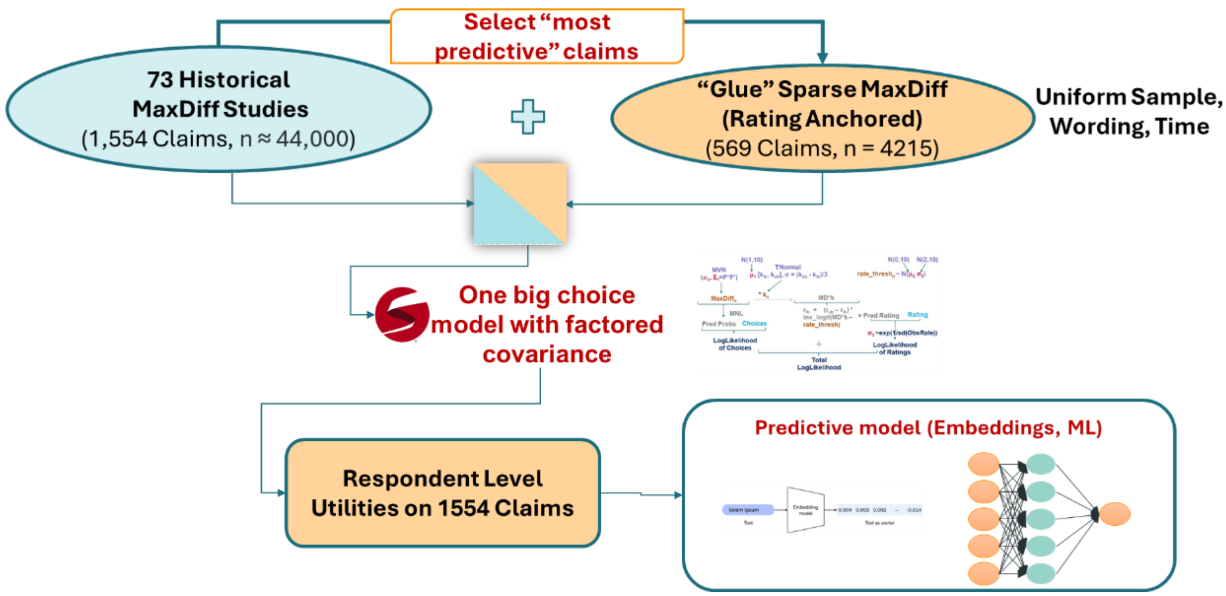
We use prompting-based predictions as benchmarks for our proposed approach. While conditioning on specific persona profiles might improve such approaches, we did not pursue this direction as it would have required additional fieldwork to obtain a representative distribution of personas. Moreover, calibration issues would remain regardless.

3. Methodology

Having identified the key challenges in both meta-analysis and predictive modeling, we now present our integrated methodology that addresses these issues systematically. Our approach addresses these challenges through a four-stage process: (1) selecting predictive claims for a glue study, (2) conducting the glue study to establish anchoring across historical studies, (3) fitting a combined model with factored covariance structure to estimate utilities for all claims simultaneously, and (4) developing language model-based predictions for new claims. Figure 2 provides an overview of the different steps.

The core challenge stems from the lack of anchoring across our 73 historical MaxDiff studies—each study produces only relative preference scores that cannot be directly compared. To create a unified preference scale across all 1,554 claims, we need a bridging mechanism.

Figure 2



Our solution is to conduct a new "glue study" that tests a strategic subset of claims from the historical studies under consistent conditions with proper anchoring. This glue study serves as the common reference point that enables us to place all historical claims on the same absolute scale.

3.1 Stage 1: Selecting Predictive Claims for a Glue Study

We first identify a subset of claims that can predict preferences for the remaining claims. For each historical MaxDiff study, we employ a conditional distribution approach based on multivariate normal (MVN) assumptions.

Given respondent utilities following MVN distribution with mean α and covariance Σ, partitioned as:

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

The conditional expectation of utilities in subset 2 given observed utilities x in subset 1 is:

$$E[U_2|U_1 = x] = \alpha_2 + \Sigma_{21}\Sigma_{11}^{-1}(x - \alpha_1)$$

We develop an algorithm to select predictive subsets:

1. Exhaustively search all 3-item combinations to find the best initial triplet
2. Sequentially add items that maximize prediction accuracy for remaining items
3. Continue until reaching desired subset size (typically 8 items per study)

This approach will seek to find subsets of items that:

1. Cannot be predicted from those already selected
2. Help predict other items
3. Have more variance

3.2 Stage 2: Conducting the Glue Study

We fielded a new anchored sparse MaxDiff study where we recruited a total of 4,215 respondents, applying uniform screening criteria to ensure that claim evaluations were consistent across the entire sample. The study itself comprised 569 claims, which were strategically selected from the 73 historical studies using our predictive subset algorithm. Each respondent was asked to evaluate 64 claims, with the assignments determined by a balanced incomplete block design. This approach provided sufficient coverage of all claims while keeping the individual respondent burden manageable.

To establish absolute preference levels, we had respondents rate how each claim would impact their likelihood to purchase. We used a 5-point rating scale ranging from "Makes me much less likely" to "Makes me much more likely." Each respondent rated a subset of items, chosen from across those items identified by the respondent as best, worst, and not chosen.

The survey was implemented using Sawtooth Software, with custom design generation to accommodate both the large number of claims and the complex blocking requirements inherent to the study design.

3.3 Stage 3: Combined Model with Factored Covariance

We develop a hierarchical Bayesian model that simultaneously estimates utilities for all 1,554 claims by combining the glue study with historical data. The model incorporates two key likelihood components mostly following the approach of Lattery (2019). First, the **MaxDiff likelihood** for choice data is given by standard multinomial logit:

$$P(y_{it} = j) = \frac{\exp(u_{ij})}{\sum_{k \in S_{it}} \exp(u_{ik})}$$

Second, the **rating likelihood** for anchor data assumes normally distributed errors of the difference between observed rating and scaled utility:

$$r_{ij} \sim N(\mu_{ij}, \sigma_r^2)$$

where the predicted rating μ is derived from utilities through the following steps:

- Scale by a respondent-specific factor: $k_i \in (0,2)$
- Transform via inverse logit of scaled utility minus respondent threshold θ_i:

$$p = \frac{1}{1 + \exp(-(k_i \times u - \theta_i))}$$

- Rescale to a range [.5, 5.5] corresponding to the 1–5 rating scale: $r = 0.5 + 5 \times p$

A major challenge in this modeling task is the high dimensionality of the covariance structure: with 1,554 claims, a full covariance matrix would require estimating over 2.4 million parameters. To address this, we use a factored covariance approach based on eigendecomposition.

Specifically, we first estimated the anchored MaxDiff with rating model above using just the glue study and its 569 claims. This gave us respondent level utilities for 569 claims. We then

1. Computed the eigendecomposition of the 569×569 glue study covariance: Σ=UDU'
2. Retained the first 50 factors that explain over 98% of the variance: $F = U_{[:,1:50]} \times \sqrt{D_{[1:50]}}$
3. Standardized the factors to obtain Fz with unit standard deviation per column
4. Modeled the full covariance as: $\Sigma = (F_z \times \text{diag}(\sigma_F))(F_z \times \text{diag}(\sigma_F))'$

This factored covariance significantly reduces the parameters in the covariance. But it is also crucial to note that we estimated respondent level utilities on the 50 factors. Each respondent's utility for each claim is a linear combination of those 50 factors. This linear combination is determined by a 50x1554 matrix of factor loadings that is the same for all respondents.

We stacked the Glue Study and 73 historical studies together creating a large MaxDiff with 1,554 claims and 48,547 respondents. Fixed factor loadings were specified for the 569 glue study claims based on the eigendecomposition results, while loadings for the remaining 985 claims were estimated from the data using standard normal priors. The factored covariance model was estimated in Stan using multi-threaded code on an AWS r7g.16xlarge instance with 64 cores and 512GB RAM.

Convergence diagnostics indicated excellent mixing, with a mean R-hat of 1.05 across all parameters. In this study, we implemented a two-stage approach, first determining the number of factors and some of the loadings based on one study, and then estimating addition loadings. Since this study, we have generalized this factored covariance approach to a single stage. This was implemented in the best paper at Sawtooth Software 2025 by Joris van Gool and Peter Li. In that study, they estimated a large MaxDiff using a single stage factored covariance matrix.

3.4 Stage 4: Language Model-Based Prediction

The goal of the predictive model is to predict the "beat average probability" (BAP) of any claim. BAP represents the probability that claim k is preferred over an average claim (utility = 0) and serves as a commonly used metric for discussing claims performance. It can be easily computed from utility estimates using the following formula:

$$BAP_k = \frac{\exp(\hat{u}_k)}{\exp(\hat{u}_k) + 1}$$

Our objective is to understand aggregate (market-level) BAP, which provides us with 1,554 observations of claims text paired with their corresponding BAP scores. To make this regression problem of BAP ~ text tractable, we transform the textual claims into numerical representations through embeddings. Specifically, we use OpenAI's text-embedding-3-large model to create 3,072-dimensional vector representations of each claim, where similar claims occupy similar positions in this high-dimensional space. These embeddings serve as feature inputs for our machine learning model, which solves the regression problem BAP $\sim f(\text{embedding})$ while maintaining the flexibility to capture complex interactions between linguistic features.

We employ AutoGluon (Erickson et al., 2020) for automated machine learning (AutoML) to optimize model architecture and avoid overfitting. To ensure robust performance estimates, we implement 5-fold cross-validation with an 80/20 train/test split throughout the model selection process. AutoGluon systematically evaluates multiple machine learning architectures.

3.5 Benchmarks and Evaluation

We compare our approach against several benchmark models encompassing both embedding-based and prompt-based methodologies. The embedding-based benchmark employs a *k*-nearest neighbors approach, calculating the average BAP score of the *k* most similar claims based on embedding distance. This method leverages the embedding model's ability to position semantically similar claims in proximity within the high-dimensional space.

The prompt-based approaches include both zero-shot and few-shot learning strategies. In the few-shot approach, we provide the language model with example claims and their corresponding scores, where examples are selected based on their similarity to the target claim. This comprehensive evaluation framework encompasses four distinct methodologies:

1. **Zero-shot prompting:** Direct prediction using GPT-4 (specifically, GPT-4-Turbo-2024-04-09) without contextual examples
2. **Few-shot learning (FSL):** Prompting with carefully selected example claims and their scores
3. **K-nearest neighbors:** Average BAP of *k* most similar claims determined by embedding distance
4. **Embedding + AutoML:** Neural network regression trained on text embeddings

To ensure robust and unbiased performance estimates, we implement 5-fold cross-validation with an 80/20 train/test split for all comparative methods. This approach provides reliable estimates of out-of-sample performance while preventing overfitting during model selection and hyperparameter optimization.

We assess model performance using three complementary evaluation metrics:

- **R^2:** Quantifies the proportion of variance in actual BAP scores explained by model predictions. Higher values indicate superior absolute score prediction accuracy.
- **Mean Absolute Error (MAE):** Represents the average absolute difference between predicted and actual BAP scores, expressed in percentage points. Lower values signify more precise predictions.
- **Pearson's r:** Measures the linear correlation between predicted and actual scores, capturing the model's ability to correctly rank claims independent of absolute score precision.

4. RESULTS

4.1 Integrated Choice Model

The combined model successfully integrated all 73 historical studies into a unified framework. The model preserved relative utilities within each historical study, maintaining the integrity of original preference orderings while establishing an absolute scale across all 1,554 claims. This enables direct comparison of claims from different studies, something that was previously impossible due to the separate scaling of individual studies.

The factored covariance approach proved highly effective, capturing 96.3% of the variance using only 50 factors. This demonstrates efficient dimensionality reduction while maintaining the essential structure of consumer preferences across the full claim set. The computational benefits were substantial, with computation time reduced by 67% compared to a full covariance model. This reduction made the analysis feasible within reasonable timeframes while preserving the statistical rigor necessary for accurate preference modeling.

4.2 Predictive Model

Table 1 shows cross-validation results for predicting beat-average probability.

Table 1

Method	R^2	MAE	Pearson's *r*
Embedding & AutoML	81.4%	6.07	0.904
Prompting & FSL	71.4%	8.29	0.879
Embedding & k-nearest neighbors	69.6%	8.15	0.849
Prompting 0-shot	-128.5%	26.69	0.551

The embedding-based AutoML approach significantly outperforms other methods, achieving:

- An R^2 of 81.4%, which explains the vast majority of variance in claim performance across the test set.
- A mean absolute error of 6.07 percentage points demonstrates high accuracy in predicting beat-average probabilities.
- A strong rank-order correlation ($r = 0.904$) indicates the model correctly identifies which claims will perform better than others.

The few-shot learning approach demonstrates competitive performance, achieving an R^2 of 71.4% and maintaining strong rank-order correlation ($r = 0.879$). This method benefits from providing the language model with contextual examples, enabling it to better calibrate its predictions against known performance benchmarks. The careful selection of similar claims as examples helps the model understand the nuanced factors that drive consumer preference within specific semantic neighborhoods.

The *k*-nearest neighbors baseline, while simpler in approach, still captures meaningful patterns with an R^2 of 69.6%. This method's effectiveness validates the quality of the embedding space, demonstrating that semantically similar claims do tend to have similar performance characteristics. However, the approach is limited by its assumption that local similarity in embedding space directly translates to performance similarity, missing the complex non-linear relationships that the neural network captures.

The zero-shot prompting approach shows poor performance with a negative R^2 of -128.5%, indicating predictions worse than simply using the mean BAP score. This failure highlights the challenge of absolute score calibration without contextual anchoring. While the method maintains some ability to rank claims ($r = 0.551$), it struggles significantly with the precise numerical prediction task, emphasizing the importance of either training data or carefully selected examples for accurate performance estimation.

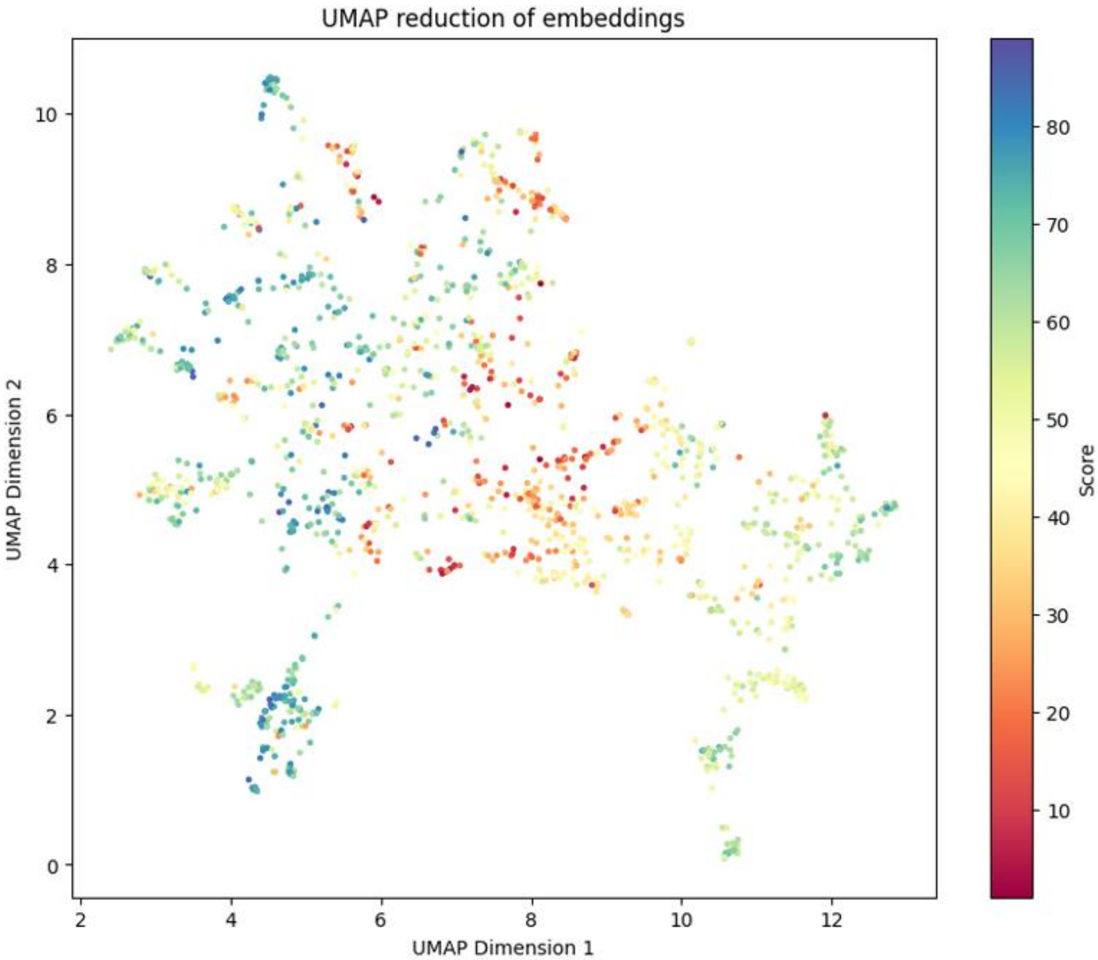
4.3 Embedding Analysis

The UMAP visualization reveals the underlying structure that enables our prediction model's effectiveness. Figure 2 presents a two-dimensional projection of the high-dimensional embedding space, where each point represents a claim colored by its beat-average probability score. The visualization demonstrates that while BAP scores do not correlate with any single global dimension, distinct clusters of high- and low-performing claims emerge throughout the space.

This clustering pattern is crucial to understanding model performance. Rather than performance being determined by position along major axes, successful claims form localized neighborhoods in the semantic space. These performance-homogeneous regions suggest that claims with similar linguistic and semantic properties tend to achieve comparable consumer preference outcomes. Our machine learning model leverages these higher-dimensional neighborhood structures to make accurate predictions, effectively learning to identify the subtle semantic patterns that distinguish high-performing claims from their lower-performing counterparts.

The embedding space thus captures not just semantic similarity, but performance-relevant similarity—a distinction that proves essential for accurate prediction. Claims that appear semantically related may perform quite differently, while claims that seem disparate may share hidden structural features that drive similar consumer responses. However, given the absence of an interpretable factorial structure, it is not straightforward to characterize clusters of low- and high-performing claims directly.

Figure 3



5. Discussion

This research contributes to the literature in several important ways. First, we advance the methodological framework for meta-analysis of discrete choice experiments by developing a systematic approach to integrate independent MaxDiff studies through strategic anchoring. The proposed glue study methodology addresses fundamental challenges in combining preference data across separate studies where traditional bridging approaches fail due to minimal item overlap. This contribution extends beyond the specific application domain and provides a generalizable framework for researchers seeking to leverage historical choice data across multiple independent studies.

Second, we contribute to the computational methods literature by demonstrating the feasibility of estimating large-scale hierarchical Bayesian choice models through factored covariance structures. The eigendecomposition approach we implement reduces computational complexity from $O(p^2)$ to $O(p \times k)$, enabling practical estimation of models with thousands of items that were previously computationally intractable. This methodological advance has implications for choice modeling applications beyond claim testing, including large-scale product positioning and assortment optimization contexts.

Third, we contribute to the emerging literature on hybrid human-AI systems in consumer research by demonstrating how preference measurement methods can be effectively combined with modern natural language processing techniques. Our approach leverages human preference data to establish ground truth while employing machine learning for pattern recognition and prediction, creating a methodological bridge between established choice modeling traditions and contemporary AI capabilities.

The implications for market research practice are substantial. The proposed methodology addresses the persistent challenge of research silos by transforming isolated preference studies into integrated knowledge systems. This represents a shift from traditional serial testing paradigms toward cumulative knowledge building in consumer research. The embedding-based prediction framework demonstrates how semantic representations can capture performance-relevant linguistic patterns, contributing to our understanding of how language features influence consumer preferences. Furthermore, the demonstrated prediction accuracy ($R^2 = 81.4\%$) suggests that automated claim evaluation may serve as a viable complement to traditional fieldwork, potentially enabling more efficient resource allocation in consumer research programs.

Several limitations should be noted. First, predictions assume new claims fall within the space defined by historical claims, though the semantic space is likely quite large. While embedding models can find similar meaning even when words differ and pay attention to linguistic patterns, model performance may degrade for radically different claim types—essentially, it works until it doesn't. Additionally, temporal effects are not explicitly modeled in the current approach, which could limit accuracy as consumer preferences evolve over time.

The current methodology focuses on aggregate preferences rather than individual differences, though extending to segment-level analysis would be straightforward. Finally, embedding dimensions lacks direct interpretability, making it difficult to understand which specific linguistic features drive predictions and limiting the actionable insights that can be drawn from the model's decision-making process.

Future work could address these limitations through several approaches. Continuous model updating with new studies would ensure predictions remain accurate as consumer preferences evolve over time. Incorporating temporal trends could help account for seasonal variations and long-term shifts in what resonates with consumers. Developing segment-specific models would enable more targeted predictions for different demographic or psychographic groups. Finally, research comparing the value of historical choice data versus re-running MaxDiff studies to inform AI and NLP-powered prediction models would provide valuable guidance on optimal resource allocation in consumer research programs.

6. Conclusion

This paper presents a comprehensive solution to a common problem in market research: transforming isolated preference studies into integrated knowledge systems with predictive capabilities. Through an empirical application involving 73 historical MaxDiff studies comprising 1,554 unique claims tested across approximately 44,000 respondents, we demonstrate how strategic anchoring through glue studies and factored covariance modeling can successfully integrate independent research efforts into a unified claims database.

Our approach addresses three fundamental methodological challenges simultaneously. First, we develop a systematic framework for integrating independent MaxDiff studies through strategically selected glue studies, enabling direct comparison across previously incomparable research efforts. Second, we advance computational methods for large-scale choice modeling through factored covariance structures, reducing computational complexity from $O(p^2)$ to $O(p \times k)$ while capturing 96.3% of preference variance with only 50 factors. Third, we demonstrate how embedding-based machine learning can achieve remarkable predictive accuracy ($R^2 = 81.4\%$) for new, untested claims, enabling instantaneous evaluation that would otherwise require weeks of traditional fieldwork.

The practical implications extend beyond methodological advancement. This work transforms serial, independent claim testing into a comprehensive prediction system, fundamentally changing how organizations can leverage historical research investments. Rather than conducting isolated studies that cannot inform future decisions, teams can now build cumulative knowledge bases that accelerate innovation cycles and improve resource allocation in product development.

As consumer research organizations accumulate ever-larger databases of historical choice data, similar approaches offer the potential to unlock significant value from past research investments while enabling more agile and informed decision-making processes. The demonstrated effectiveness of combining traditional preference measurement with modern machine learning techniques provides a roadmap for the future evolution of applied consumer research.

Jeremy Christman

Kevin Lattery

Nino Hardt

Liz Clevenger

David Hengehold

Pankaj Patil

Howard Huang

APPENDIX

Stan Model Code Excerpt

```
data {
 int<lower=1> I; // number of respondents
 int<lower=1> P; // total number of items (1554)
 int<lower=1> P_fac_load_fix; // number of factors (50)
 int<lower=1> P_load_fix; // items with fixed loadings (569)
 matrix[P_load_fix, P_fac_load_fix] F_z_fixed; // fixed factor loadings
```

```
 // ... additional data declarations
}

parameters {
 matrix[I, P_fac_load_fix] U_z; // respondent factor scores
 matrix[P - P_load_fix, P_fac_load_fix] F_z_est; // estimated loadings
 vector<lower=0>[P_fac_load_fix] sigma_F; // factor standard deviations
 // ... additional parameters
}

model {
 matrix[P, P_fac_load_fix] F_z_full;

 // Combine fixed and estimated loadings
 F_z_full[1:P_load_fix, :] = F_z_fixed;
 F_z_full[(P_load_fix+1):P, :] = F_z_est;

 // Prior on estimated loadings
 to_vector(F_z_est) ~ std_normal();

 // Compute utilities
 matrix[I, P] U = U_z * diag_post_multiply(F_z_full', sigma_F);

 // MaxDiff likelihood
 // ... choice model implementation

 // Rating anchor likelihood
 // ... rating model implementation
}
```

Example Prompts

To ensure reproducibility, we provide examples of the prompts used:

Zero-shot prompt:

```
Context: Imagine a US consumer evaluating [oral care products] based on
claims or descriptions.
Objective: Considering the context, infer how much the claim or description
makes someone buy the product, on a scale of 0-100; a score of 50 indicates a
50% chance of the description performing better than an average-performing
description.
Criteria:
- Relevance: How well the claim/descriptions addresses the needs of the
consumer.
- Utility: How much value the claim/description adds
Response: Provide the score only.
Focal Concept: [CLAIM TEXT HERE]
```

Few-shot learning prompt:

```
Context: Imagine a US consumer evaluating [oral care products] based on
claims or descriptions.
Objective: Considering the context, infer how much the claim or description
makes someone buy the product, on a scale of 0-100; a score of 50 indicates a
50% chance of the description performing better than an average-performing
description.
Criteria:
- Relevance: How well the claim/descriptions addresses the needs of the
consumer.
- Utility: How much value the claim/description adds
Response: Provide the score only.
Examples:
- [Example 1]: 72
- [Example 2]: 58
- [Example 3]: 65
- [Example 4]: 81
- [Example 5]: 61
Focal Concept: [CLAIM TEXT HERE]
```

Computational Resources

All computations were performed on AWS:

- Glue study estimation was conducted on a c7g.16xlarge instance with 64 vCPUs and 128 GB RAM to handle the initial model fitting.
- The combined model required an r7g.16xlarge instance with 64 vCPUs and 512 GB RAM due to the memory demands of the 1,554-item covariance matrix.
- The prediction service runs efficiently on a t3.medium instance with just 2 vCPUs and 4 GB RAM, demonstrating the lightweight nature of the deployed model.

Total computational cost: approximately $2,500 USD

REFERENCES

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., & Smola, A. (2020). AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. *arXiv Preprint arXiv:2003.06505.*

Lattery, K. (2019). Anchoring Maximum Difference Scaling Against a Threshold—Dual Response and Direct Binary Responses. *Sawtooth Software Conference Proceedings.*

Stan Development Team (2024). Stan Reference Manual, 2.36. https://mc-stan.org

Y. Gao, D. Lee, G. Burtch and S. Fazelpour (2025). Take caution in using LLMs as human surrogates, PNAS 122 (24).

AI AT MICROSOFT: ENHANCING CONJOINT AND SCALING IN-DEPTH INTERVIEWS

DANIEL PENNEY
SUHASINI SANYAL SAXTON
ANANYA RAMJE
MICROSOFT
EVAN DIMARZIO
PSB INSIGHTS

I. ABSTRACT

In the evolving landscape of market research, understanding customer preferences and decision-making processes is critical for businesses seeking to remain competitive. One of the most effective methodologies in this domain is conjoint analysis, a key part of any researcher's toolkit that can quantify how customers value different attributes of a product or service by simulating real-world trade-offs and is considered the gold standard for exploring customer decision-making.

However, traditional approaches to conjoint analysis face significant challenges. Holding respondent engagement is difficult in long exercises or exercises with many moving parts, which can lead respondents to take shortcuts to finish the exercise. This in turn can reduce the accuracy of the data and impact its alignment with real-world customer behavior. The integration of artificial intelligence (AI) into market research presents an opportunity to address these challenges. AI tools have the potential to enhance respondent engagement through more personable engagement, which in theory can improve data quality and streamline research processes.

This paper covers 3 separate experiments, exploring the application of AI in enhancing conjoint analysis in experiments 1 and 2, and scaling in-depth interviews (IDIs) in experiment 3. The first experiment explores the impact when open-ended questions are paired with AI-generated probes within the conjoint exercise, while the second expands this to test the impact of having voice responses to the open-ends. In the third experiment, AI-moderated interviewing is evaluated as a viable alternative to human moderation (in part, to potentially scale in-depth interviews, augmenting human IDIs with AI-interviews). Together these three experiments evaluate the impact of conversational AI on respondent experience, data quality, and research efficiency. The findings cover benefits, pitfalls and considerations from the experiments, and provide actionable insights into the integration of AI into these research methodologies, offering a pathway toward achieving quantitative insights with qualitative depth.

II. LITERATURE REVIEW

Conjoint analysis has been a foundational tool in market research since its introduction by Green and Rao (1971). It allows researchers to quantify consumer preferences by simulating real-world trade-offs, with discrete choice modeling (DCM) emerging as a dominant technique for capturing decision-making processes (Louviere et al., 2000). However, traditional conjoint methods often face challenges such as respondent fatigue and limited ability to capture the underlying motivations behind choices (Orme, 2014). These limitations hinder the richness of insights and alignment with real-world behavior. Findings pointed to methods that could enhance the insights from conjoint, such as “replays” of conjoint with qualitative interviews conducted after the exercise (Poynter, 1999).

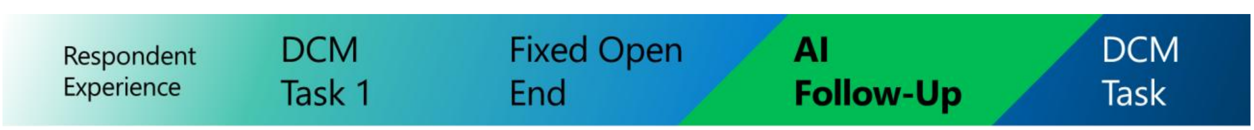
In-depth interviews (IDIs), on the other hand, provide qualitative depth by exploring the “why” behind consumer decisions (McCracken, 1988). While IDIs excel at uncovering nuanced insights, they are resource-intensive and difficult to scale, making them impractical for large or time-sensitive studies (Boyce and Neale, 2006). The need for approaches that combine the scalability of conjoint analysis with the depth of IDIs has driven interest in AI-enhanced methodologies.

In the context of response modalities, Hirschberg and Manning (2015) noted that voice-based responses often feel more engaging and intuitive for participants but pose challenges related to transcription accuracy and recruitment. Later research by Karty (2024) indicated high potential for voice research to enhance responses, motivating deeper exploration into the applications of voice. Nguyen et al. (2022) further highlighted the potential of AI-moderated interviews, finding that they produced results consistent with human moderation in structured scenarios, though hybrid models combining AI and human input may offer the best balance of scalability and depth.

III. Methodology—Experiment 1: Opens + AI Probes in Conjoint

Overview of the Experiment

This experiment was conducted to evaluate the potential of open-ended questions and artificial intelligence (AI) probes to enhance conjoint analysis. The hypothesis is this will improve the quality and depth of open-ended responses and result in better quality of the model by having respondents reflect more deeply on their choices. Open-ended questions will be structured to better answer the “why” behind the respondent decision. Traditional conjoint analysis uses choice-based tasks, which, while effective for quantitative modeling, often fail to capture nuanced qualitative insights. By integrating a series of strategically placed open-ended questions followed by AI-driven probe questions into the survey design, this study aimed to determine whether these enhancements could increase respondent engagement, generate richer data, and improve the alignment of results with real-world customer behavior. This is especially important when looking for more considered decision-making rather than instinctive or reactive decision-making, where slowing down respondents would not necessarily benefit the exercise as much. In this specific example, decision making is done by commercial IT professionals that impact large companies, and so a slower, more considerate decision is desirable.



Experimental Design

The experiment utilized a between-subjects design, dividing respondents into two distinct groups (referred to as "cells") to compare traditional conjoint analysis with an AI-enhanced approach. Respondents in these studies are IT decision-makers at organizations larger than 300 employees. The experimental conditions were as follows:

- Cell A: Conjoint Without Open-Ended Questions

This condition served as the control group. Respondents completed a standard discrete choice modeling (DCM) exercise consisting of ten tasks without any open-ended questions. The sample size for this cell was 800 respondents.

- Cell B: Conjoint With Open-Ended Questions and AI Probes

In this condition, respondents completed a modified DCM task that included eight tasks, three of which were followed by open-ended questions. Each open-ended question included an AI-powered probe, prompting respondents to elaborate further on their answers and encouraging more detailed and thoughtful responses. The sample size for this cell was 175 respondents.

The open-ends in Cell B were strategically placed through the conjoint exercise after tasks 1, 3, and 5 to provide "checkpoints" for respondents, gathering information on the thinking behind the decision and an understanding of where the respondent places value. To keep the overall length of interview comparable, respondents in Cell B were only shown 8 conjoint tasks instead of the 10 that were shown to Cell A. Both conjoint exercises included the same number of attributes (17) and levels (2–5 per attribute).

Data Collection and Analysis

Data collection involved capturing both quantitative and qualitative inputs from respondents across the two experimental cells. The primary outcomes measured in this experiment included:

- **Response Verbosity:** The length and detail of open-ended responses provided by participants.
- **Model Fit:** The percentage of correctly predicted tasks when applying the model to the training data, serving as a measure of how well the conjoint model aligned with training data.
- **Respondent Experience:** Indicators of participant engagement and the overall quality of their interaction with the survey.

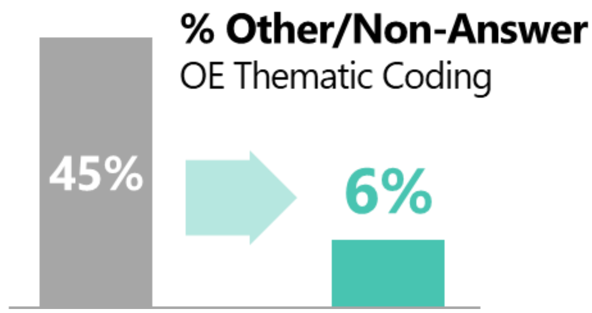
For Cell B, open-ended responses were analyzed using thematic coding to assess their richness and information. Comparisons were made between the two cells to identify differences in respondent engagement, the depth of insights generated, and the overall quality of the data. Conjoint responses were analyzed using the CBC HB standalone product created by Sawtooth Software. Open-ended responses with probes were analyzed using PSB's proprietary thematic coding tool, with open-ended themes to be explored as covariates in the modeling.

IV. Results—Experiment 1: Opens + AI Probes in Conjoint

Comparison of Results Between Cells

The results of the first experiment revealed notable differences between the two experimental conditions.

Key Findings



The inclusion of AI probes in Cell B led to a substantial increase in the length and detail of open-ended responses. On average, responses to the first open-end in Cell B were 57% longer when incorporating the AI-probed response. Beyond simply increasing the word count, the increase in verbosity translated into deeper insights as well, as the percent of "Other/None" responses drawn from AI-coded themes was reduced from 45% when including just the initial response, to 6% when including both the initial and AI-probed responses.

The DCM model in Cell B demonstrated a stronger fit to training data compared to Cell A. Specifically:

1. Cell A (No Open-Ended Questions): To ensure comparability to Cell B, 8 of 10 tasks were randomly selected for analysis. With this condition, the model achieved a fit of 90% to the training data.
2. Cell B (with opens and AI probes): The model achieved a fit of 96%, reflecting a marked improvement in the ability to fit the training data. When open-ended responses from Cell B were used as covariates in the model, the fit remained at 96%.

DCM Quality Metrics	Fit to Training Data
DCM with No Open-Ends	90%
DCM with Opens and AI probes	**96%**
DCM with Opens (+AI probes) as Covariates	**96%**

Because the sample and design were consistent between the two studies, this increased model fit is likely a result of better respondent engagement, giving less noise as well as cleaner data.

Data from Cell B also showed stronger alignment with real-world customer behavior, particularly in terms of product mix preferences and what we believed about price sensitivity in this case. For Cell B, share of preference for the primary product, a lower cost option, showed closer alignment with the actual market share. The increased price sensitivity in Cell B was also more in line with the expected behavior of the marketplace. By eliciting more detailed responses, the AI-enhanced approach provided insights that were more reflective of actual market dynamics.

Summary of Results

The findings from Experiment 1 demonstrate that incorporating AI probes into conjoint analysis can enhance the quality of data collected. By increasing verbosity, improving model fit, and achieving better alignment with real-world behaviors, the use of AI probes addresses key limitations of traditional conjoint analysis. Beyond strictly statistical significance, the incorporation of open-ends and AI probes gave deep insight into the customer's "why" behind their decision. Open-ended data from the conjoint modeling proved helpful in illustrating to project stakeholders why customers were making their choices and bringing to life the voice of the customer. The increased depth of insights and benefits to the modeling gave strong evidence that incorporating open-ended questions and AI-guided follow-ups should be considered a valuable addition to the conjoint toolset.

V. Discussion—Experiment 1: Opens + AI Probes in Conjoint

Implications and Benefits of Using AI in Conjoint Analysis

The findings from Experiment 1 highlight the transformative potential of artificial intelligence (AI) in enhancing conjoint analysis by addressing key limitations of traditional methods. One of the most significant contributions of AI-driven probes is their ability to elicit richer, contextually nuanced responses that complement structured data. By encouraging respondents to elaborate on their choices, AI probes not only improve response verbosity but also foster greater engagement with the survey process. This enhanced interaction translates into deeper insights, enabling researchers to better understand the "why" behind consumer decisions.

The improved fit of discrete choice modeling (DCM) in Cell B underscores how AI can indirectly impact the accuracy and reliability of quantitative models. The reduction in non-informative responses—from 45% in the first response for Cell B to just 6% when incorporating the AI-probed OE—further demonstrates the role of AI in guiding participants toward providing actionable feedback. These improvements strengthen the model's sensitivity to critical factors such as price elasticity and product mix preferences, making it more reflective of real-world customer behavior.

Beyond data collection, the integration of conversational AI bridges the gap between quantitative and qualitative research. Traditional conjoint analysis often sacrifices contextual detail for scalability, but AI probes allow researchers to capture qualitative nuance without compromising the structured nature of the survey. This hybrid approach enriches the utility of conjoint analysis for predictive modeling and decision-making, offering insights that are both statistically robust and qualitatively meaningful. As demonstrated in this study, AI tools represent a powerful innovation for advancing market research methodologies and improving the depth and reliability of insights.

Pitfalls and Considerations

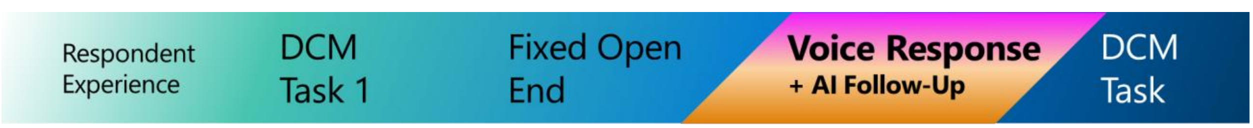
While the results of Experiment 1 highlight the benefits of using AI in conjoint analysis, there are important considerations and potential pitfalls that must be addressed. The integration of AI probes requires careful calibration to ensure that follow-up prompts are relevant and do not introduce bias into respondents' answers. Poorly designed or overly intrusive probes could risk frustrating participants, leading to disengagement or lower-quality responses. In addition, later research showed the importance of varying the questions asked. When asked the same or too similar questions, respondents grew frustrated as they felt they had already responded to that question (and they had). Finally, while AI probes improve the depth of open-ended responses, their effectiveness depends on the context and audience. Certain populations may find AI-generated follow-ups less intuitive or engaging, potentially impacting the generalizability of the findings. Further testing across diverse demographic and cultural groups is necessary to validate the broader applicability of this approach.

Building on the findings from Experiment 1, the second experiment examines how response modality—text versus voice—further impacts data quality and respondent experience.

VI. Methodology—Experiment 2: Voice vs. Text for OE Questions in Conjoint

Overview of the Experiment

The second experiment was designed to evaluate the impact of response modality—text-based versus voice-based—on the quality and depth of open-ended responses in conjoint surveys. While text-based responses are commonly used in market research due to their simplicity and scalability, voice-based responses have the potential to capture richer, more nuanced feedback by allowing respondents to articulate their thoughts verbally, allowing for longer and more detailed responses. This experiment sought to determine whether the inclusion of voice-based responses could simplify and enhance respondent engagement, improve the quality of open-ended data, and provide additional insights compared to the previously tested text-based input.



Experimental Design

This experiment employed a between-subjects design, dividing respondents into two cells based on the mode of response for open-ended questions. Respondents in these studies are IT decision-makers at organizations larger than 300 employees. Both groups completed a discrete choice modeling (DCM) task with identical structures, except for the modality of their open-ended responses:

- Cell X: Text-Based Responses
  Respondents in this group provided open-ended feedback using a text-input format. The survey consisted of 8 tasks, including 3 open-ended questions with AI-driven follow-up probes. The sample size for this cell was 700 respondents.
- Cell Y: Voice-Based Responses
  In this group, respondents provided open-ended feedback verbally, which was recorded and transcribed for analysis. Like Cell X, the survey included 8 tasks with 3 open-ended questions enhanced by AI follow-up probes. The planned sample size for this cell was 200 respondents.

Respondents in Cell Y were informed that they would be required to respond using voice and asked to enable their microphone early in the survey. Both cells received identical conjoint designs with the same number of attributes (12) and levels (2–4 per attribute).

Data Collection and Analysis

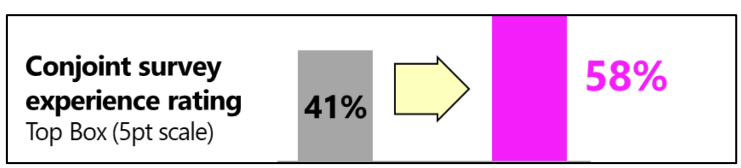
Data collection involved capturing both quantitative and qualitative inputs from respondents in each cell. For text-based responses (Cell X), open-ended answers were directly analyzed using thematic coding to evaluate verbosity and information richness. For voice-based responses (Cell Y), verbal feedback was transcribed prior to analysis, ensuring consistency in the application of thematic coding. Conjoint responses were analyzed using the CBC HB standalone product created by Sawtooth Software. The primary variables measured in this experiment were:

- **Response Length:** Measured as the average word count for text responses and transcriptions of voice responses.
- **Proportion of Information-Rich Responses:** Calculated as the percentage of open-ended responses containing actionable insights or detailed explanations, defined as covering 3 or more themes across their responses.
- **Survey Experience Ratings:** Collected through post-survey evaluations, where respondents rated their experience on a five-point scale.
- **Model Fit to Training Data:** Assessed for both cells to determine whether the inclusion of voice-based responses influenced the predictive accuracy of the conjoint model.

Comparisons between the two cells were conducted to identify differences in response quality and respondent experience, as well as to determine the feasibility of implementing voice-based responses in large-scale conjoint studies.

VII. Results—Experiment 2: Voice vs. Text for OE Questions in Conjoint

Differences in Key Metrics



The results of Experiment 2 revealed distinct differences between text-based (Cell X) and voice-based (Cell Y) responses across several key metrics. Respondents in Cell Y (voice-based responses) reported a more enjoyable survey experience compared to those in previous conjoint survey experience measures. Post-survey evaluations showed that 58% of respondents in Cell Y rated their experience as "top box" on a five-point scale, compared to 41% in an average conjoint survey. This suggests that the voice modality provided a more engaging and interactive format for participants.

Voice-based responses were significantly longer than text-based responses, with an average increase of 21% in verbosity. Thematic coding revealed that 67% of text-based responses contained 3 or more actionable insights, compared to 57% of voice-based responses. While voice responses were longer, they tended to include more conversational filler and less structured feedback, reducing their overall utility for modeling purposes. Modeled results were very similar between the two sections, with nearly identical fit to training data and model scale factors.

DCM Quality Metrics	Fit to Training Data
Cell X: DCM with Text Response	89%
Cell Y: DCM with Voice Response	**89%**

Challenges with Recruiting and Fielding for Voice-Based Responses

Despite the potential benefits of voice-based input, significant challenges were encountered during recruitment and fielding. Voice-based surveys required respondents to have access to recording devices and sufficient comfort with verbal communication. The mandatory use of voice responses contributed to higher dropout rates, as some participants reported finding the voice modality cumbersome. The result was delays and increased fielding time, and in the end only two-thirds of the intended sample size for Cell Y was successfully recruited (131 of 200).

VIII. Discussion—Experiment 2: Voice vs. Text for OE Questions in Conjoint

Implications of Findings

The findings from Experiment 2 illustrate important trade-offs between text-based and voice-based responses in conjoint surveys. While voice-based responses offer a more engaging experience for participants, in this particular experiment they did not necessarily yield richer or more actionable data compared to text-based responses. The increased verbosity observed in voice responses suggests that participants feel more comfortable elaborating verbally; however, this did not translate into greater thematic depth or clarity. Text-based responses, by contrast, tend to be more concise and structured, making them easier to analyze and integrate into modeling and reporting. From a methodological perspective, these results suggest that voice-based responses may be best suited for scenarios where respondent engagement is a priority, such as exploratory research or studies involving sensitive topics where verbal communication might elicit more candid feedback. However, for tasks requiring highly structured and actionable data, text-based responses remain a practical and reliable option.

Pitfalls and Considerations

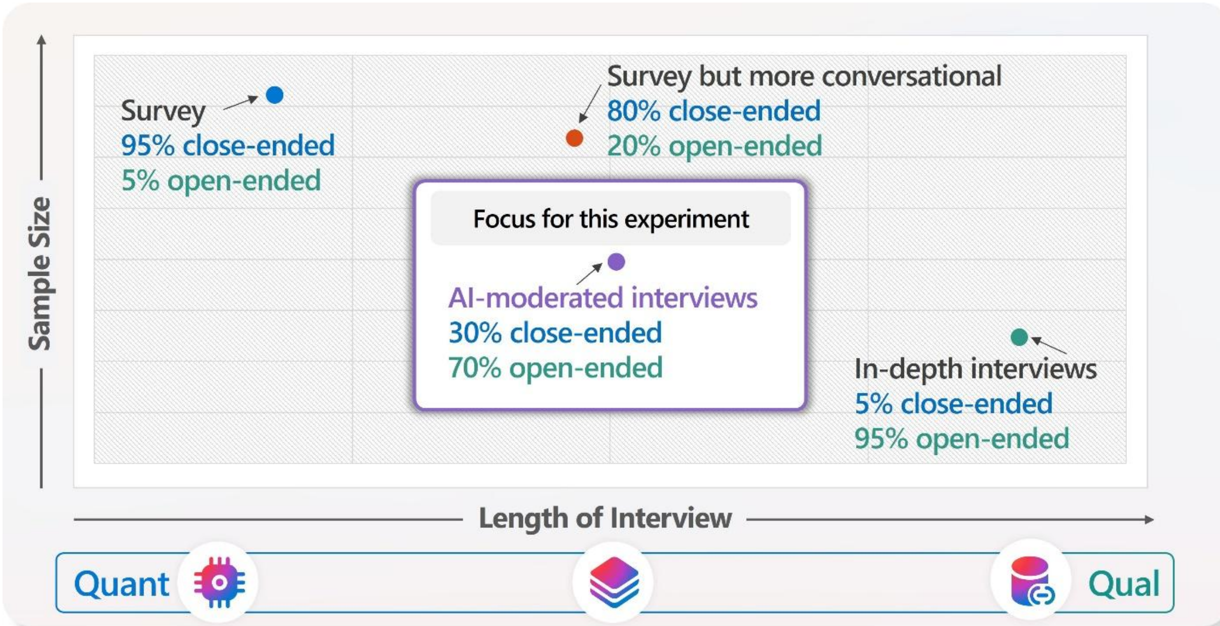
While voice-based responses offer certain advantages, their implementation poses several challenges that must be carefully managed. Recruitment difficulties were a major obstacle in this experiment, as the requirement for voice input limited the pool of eligible participants. Additionally, the increased dropout rate associated with this voice-based survey indicates that some respondents may find this modality inconvenient or unappealing, which is particularly impactful in large-scale studies with focused audiences. Fielding voice-based surveys also requires additional resources, including transcription services and quality control measures to ensure the accuracy of recorded responses. These added complexities can increase costs and extend timelines, potentially offsetting the benefits of using voice input. Increased verbosity initially looks like a win, but this did not translate into more actionable insights drawn from the responses.

Given these considerations, researchers should approach the use of voice-based responses with caution and consider offering it as an optional modality rather than a mandatory requirement. Testing and refining voice-based methods across diverse audiences and contexts will be essential to fully understand their feasibility and effectiveness in market research applications.

IX. Methodology—Experiment 3: AI-Moderated Interviews

Overview of Experiment: Scaling In-Depth Interviews with AI

As artificial intelligence continues to reshape the research landscape, Microsoft has been actively exploring how AI can enhance traditional methodologies. One of the most promising frontiers is the use of conversational AI to scale in-depth interviews (IDIs), a method traditionally limited by time, cost, and logistical constraints. In this context, conversational AI leverages natural language processing (NLP), machine learning, and generative AI to engage participants in real-time, human-like dialogue. Unlike traditional surveys or interviews, these systems can dynamically adapt questions based on responses, probe deeper, and even summarize insights on the fly. In collaboration with multiple research partners, we conducted a series of experiments to evaluate the feasibility, effectiveness, and limitations of AI-moderated interviews. This section of the paper outlines our findings and offers guidance for researchers considering this approach.



Research is broadly classified into two main categories: quantitative and qualitative, though it can encompass many other forms and methods. Quantitative research usually involves large sample sizes but provides limited nuance considering surveys are typically 15–20 mins long to maintain engagement. On the other hand, qualitative research provides deeper understanding but tends to be more expensive and time-consuming. This makes it challenging to have large sample sizes, especially in B2B or commercial research where quick answers to critical business questions are needed.

With conversational AI entering the spectrum, various new opportunities are now available that were previously inaccessible. It is now feasible to conduct 30-minute interviews with commercial audiences and scale them to larger sample sizes than what was typically achievable in qualitative research due to cost and time constraints.

We have now tested AI-moderated interviews with various audiences, including B2B decision-makers like IT leaders and developer decision-makers, gamers, and general consumers. These tests covered use cases such as concept testing, pricing, purchase journeys, and jobs-to-be-done frameworks. When the experiment was conducted in the summer of 2024, this method complemented traditional approaches. AI-moderated interviews are now being explored as a standalone method to meet research objectives, but this is still under experimentation.

Experiment Design

Our third experiment aimed to compare traditional human-moderated in-depth interviews (IDIs) with AI-moderated alternatives. Microsoft partnered with CMB to conduct our first such experiment. The objective was to determine whether conversational AI could match the depth and nuance of human-led interviews while providing greater scalability and efficiency. This experiment took place in June 2024, supplementing existing research with AI-facilitated interviews. Original research involved a human moderator who conducted a total of 30 interviews, each lasting 50 minutes. The original question guide comprised a mixture of open and closed questions. The recruitment and scheduling process for the interviews took approximately 2.5 weeks, followed by another 3 weeks for conducting them.

In this experiment, the AI interviews were designed to be more conversational than surveys but more structured than traditional qualitative interviews, consisting of approximately 30% close-ended (rating-based, single select, etc.) and 70% open-ended questions. This ratio could be adjusted to suit specific research objectives. We shortened our original research question guide to 30 minutes by cutting introductory sections and some open-ended exploratory questions. A human researcher prepared a discussion guide and programmed questions into an AI platform.

Data Collection and Analysis

Each AI interview was scheduled for 30 minutes, but the actual duration varied among participants. Following each AI interview, there was a 10-minute human conversation with participants to gather feedback on this new approach. The recruitment method, criteria, and panel partner remained consistent across both methods. Respondents were recruited and completed interviews asynchronously. The recruitment took longer due to follow-up interviews, lasting three weeks. As this approach scales, sending AI interview invites like surveys can enable faster completion within days.

In this experiment, participants engaged with the AI through video, while the AI provided text-based responses for the participants to read from the screen before recording their replies. Depending on the specific requirements of the use case, audio or text-only modes can also be utilized.

Three criteria were used to evaluate the feasibility of AI-moderated interviews:

1. Participant satisfaction with the experience and likelihood of repeating it.
2. Alignment of results with those obtained from human-conducted interviews.
3. Appropriateness of this approach for various scenarios.

X. Results—Experiment 3: AI-Moderated Interviews

Respondent Engagement and Experience

Participants found AI-moderated interviews engaging and were pleasantly surprised by their effectiveness. Most appreciated the logical flow and structure. In our initial experiment, all expressed willingness to repeat the experience, encouraging us to scale this approach with senior commercial decision-makers. The asynchronous format allowed thoughtful responses, and the absence of a human moderator created a non-judgmental setting. This approach requires people to become familiar with technology, as there are no humans to build rapport with respondents.

Insight Alignment with Human-Moderated Interviews

The findings from AI-moderated interviews were consistent with those from human-led sessions. Greater price sensitivity was observed in the AI-moderated group. When presented with three product options, participants in the AI-moderated group more frequently selected the cheapest option as their preference. There could be various reasons for this behavior, including the absence of social desirability bias or differences in sample composition. Additional experiments are needed to validate or refute these findings. Although the AI did not surpass human moderators in generating insights, it can deliver comparable outcomes with substantially reduced time and cost as operations scale.

Use Case Suitability for AI-Moderated Interviews

AI-moderated interviews are suitable for scenarios requiring a combination of quantitative structure and qualitative detail. This format works well with a structured question guide that includes both closed- and open-ended questions, where detailed information on specific topics is needed. It can be an effective method when there is a need to adjust the research midstream. This approach is generally used when the research topic is moderately exploratory but not entirely unstructured. Additional exploration is conducted only for questions requiring further probing or where respondents provide incomplete information. Lastly, evaluate whether this approach can work with your target audience. For example, tech savvy commercial audiences like the IT decision-makers or Developers we interviewed found this format engaging but it may not work with consumers who may be less familiar with AI.

XI. Discussion—Experiment 3: AI-Moderated interviews

Benefits

This approach combines the rigor of surveys with the depth of qualitative feedback, particularly valuable when constrained by time or cost limitations. The primary benefit of this method is its ability to be rapidly deployed across extensive and diverse samples. Although it does not imply an absence of cost savings compared to traditional human-moderated methods, significant cost reductions are realized when the approach is scaled, involving numerous AI-moderated interviews. The principal factor in cost reduction is the elimination of human moderation fees, which also simplifies scheduling logistics, thereby enhancing efficiency. Depending on the AI interview platform utilized, aside from the initial licensing fee, the incremental cost of conducting additional interviews remains relatively low. It is important to note that costs may vary based on factors such as recruitment methodology, interview length (30 vs. 45 minutes), the extent to which the discussion is open- or closed-ended and so on.

Pitfalls

AI has limitations such as lacking the ability to connect disparate insights or read emotional cues like humans do. For instance, when a respondent provides contradictory information during a conversation, a human can seek clarification, but AI may not always be able to do so. AI cannot adjust dynamically as a human can, so probing or interview flow strictly follows the inputs provided to the AI. Unlike traditional research where interviews are moderated live, AI-based interviews might require a different approach to reporting since no person is listening and taking notes in real time. Nevertheless, many AI interview platforms offer some level of auto-generated reporting capabilities. Additionally, it is important to address privacy concerns upfront by being transparent about AI usage and data handling to maintain trust. Finally, expect some variability in interview length due to unique respondent responses. This method requires careful testing to ensure consistency.

Considerations

This approach is suitable for mixed methods with both open and closed questions. A fully exploratory conversation may be better managed by traditional human-moderation, as controlling the direction of conversation can be more challenging when questions are not as specific. However, this could change as technology advances. Additionally, respondents record their responses as audio/video after reading the question on screen, allowing them time to reflect before recording. This tends to produce system 2 responses rather than immediate reactions, which is beneficial for research requiring deliberate and thoughtful responses. Lastly, this setup may encourage more candid responses on sensitive topics, such as pricing research where overstatement often occurs.

XII. Recommendations

Best Practices for Integrating AI in Research

The findings from this study highlight the significant potential of artificial intelligence (AI) to enhance market research methodologies, particularly in conjoint analysis and in-depth interviews (IDIs). To maximize the benefits of AI while addressing its limitations, the following best practices are recommended:

1. **When to Use OE Questions and AI Probes in Conjoint Studies**
   AI probes should be incorporated into conjoint studies when there is a need to enrich choice responses without compromising the structured nature of the survey. They are particularly effective in scenarios where qualitative nuance is critical for understanding customer preferences, such as concept testing or pricing sensitivity studies. However, researchers must ensure that AI prompts are carefully designed to avoid introducing bias or leading respondents. This method is likely only appropriate in certain circumstances when you want more reflective opinions rather than instinctive or reflexive decision making.
2. **Offering Voice as an Option Rather Than a Requirement**
   While voice-based responses can improve respondent engagement and provide richer feedback in certain contexts, they also present logistical challenges, including recruitment difficulties and increased dropout rates. To address these issues, voice input should be offered as an optional modality rather than a mandatory requirement. This approach allows participants to choose their preferred response format, thereby improving accessibility and minimizing barriers to participation.
3. **Testing and Refining AI-Moderated Interview Formats**
   AI-moderated interviews have shown promise as a scalable alternative or supplement to traditional human-moderated interviews, but further refinement is necessary to optimize their effectiveness. It will often be best to have a combination of interviews driven by human moderators and AI, to achieve greater scale while ensuring depth and nuance. Audience should also be taken into consideration, as there may be a biased subset of audience members willing to participate in AI-moderated interviews. Researchers should test AI moderation across diverse topics and audiences to identify the most appropriate use cases. Additionally, iterative improvements to conversational AI systems—such as

enhancing their ability to interpret nuanced responses and maintain natural dialogue—will be essential for achieving consistent and reliable results.

XIII. Conclusion

Summary of Key Findings and Implications

This study demonstrates the transformative potential of AI in enhancing conjoint analysis and scaling in-depth interviews. Incorporating open-ends followed by AI probes into conjoint surveys showed the potential for significantly improving the quality and depth of responses, resulting in better model fit and stronger alignment with real-world customer behavior. Similarly, AI-moderated interviews emerged as a valuable complement to human moderation, offering scalability and cost-efficiency while maintaining consistency in findings. These advancements highlight the ability of AI to bridge the gap between quantitative and qualitative research, enabling richer and more actionable insights. The integration of AI into market research methodologies represents a paradigm shift in how data is collected, analyzed, and applied. AI tools not only enhance respondent engagement and data quality but also enable researchers to scale traditionally resource-intensive methods like IDIs. As AI continues to evolve, its role in shaping research methodologies will expand, providing new opportunities to explore customer behavior with greater precision and agility.

As AI continues to evolve, its role in market research will undoubtedly expand, offering new opportunities to explore consumer behavior with unprecedented depth and efficiency. This progress must be accompanied by rigorous testing, thoughtful implementation, and an ongoing commitment to balancing innovation with methodological rigor. By leveraging the insights from this study, researchers and practitioners can confidently integrate AI into their work, driving smarter decision-making and uncovering deeper truths about the complexities of consumer preferences.

Daniel Penney

Suhasini Sanyal Saxton

Ananya Ramje

Evan DiMarzio

XIV. References

Boyce, C., & Neale, P. (2006). Conducting in-depth interviews: A guide for designing and conducting in-depth interviews for evaluation input. Pathfinder International.

Green, P. E., & Rao, V. R. (1971). Conjoint measurement for quantifying judgmental data. Journal of Marketing Research, 8(3), 355–363.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing for conversational AI. Communications of the ACM, 59(11), 92–101.

Karty, K. (2024). Humanizing Surveys and Enhancing Depth of Insights Using LLMs. Sawtooth Software Research Paper Series.

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

McCracken, G. (1988). The long interview. Sage Publications.

Nguyen, T., Patel, S., & Rao, K. (2022). Evaluating AI-moderated interviews: A comparative study with human moderators. International Journal of Market Research, 64(1), 45–62.

Orme, B. K. (2014). Getting started with conjoint analysis: Strategies for product design and pricing research (3rd ed.). Research Publishers LLC.

Poynter, R. (1999). But Why? Putting the Understanding into Conjoint. Sawtooth Software Research Paper Series.

Like to Get to Know You Well: Strategies for Streamlining Variables to Create Actionable and Engaging Segmentation Models

Tracey Di Lascio- Martinuk
Bose Corporation

Introduction

This case study offers a high-level look at our Consumer Research and Insights team's process for building an insightful and actionable consumer segmentation model to illuminate consumer types within the premium audio market. While there is no shortage of technical detail involved in building a segmentation model, the most critical work often happens before a single line of code is written, so this paper focuses on that early stage: the business conversations, stakeholder engagement, and strategic decisions that shaped the model before we started on the technical side of the work.

By investing time upfront to understand our stakeholders' goals, use cases, and success criteria, we engaged the right people in the room, aligned expectations, and narrowed down our dimensions of interest before modeling even began. That foundation made variable selection and model building faster and easier, and ultimately led to a model that our teams trusted and were eager to use.

First, we identified our key stakeholders and engaged in conversations to understand their needs, wants, and what would make them feel good about adopting our model. As we discussed what they wanted to know about premium audio consumers, we narrowed down the dimensions we would use to define the segments, which helped us reign in the list of variables. Next, we designed and executed the survey. Once we had our data, we refined the variable list yet again before building, refining, and selecting our model. Then finally, we put the model into action as our stakeholders envisioned. The following sections describe each of these steps in more detail.

Identify Key Stakeholders

To build a segmentation model that our teams would find helpful and enjoyable to use, our first step was to identify our key stakeholders and understand what "actionable" looked like to them. We used the following questions to guide this first part of the process:

- **Who will use the segmentation?** We identified the organizations within the business that were the most excited for the segmentation, and where we thought it could add the most value as a decision driver. We invited decision makers from those functions into the conversation from the start.
- **Who are the primary and secondary stakeholders?** Even after narrowing the list, we still had many interested stakeholders, each of whom had their own use cases. We worked to clarify expectations early, to be sure we were all aligned to the planned outcome.
- **How will they use the segments?** We asked our stakeholders for clear examples of the decisions they hoped the segmentation would inform. When needed, we pushed for

specificity and asked them to distinguish between "must-haves" and "nice-to-haves." This helped us prioritize use cases if tradeoffs became necessary.

In our case, the model was intended to inform decisions in product development, portfolio development, and go-to-market strategy. For example, we wanted to understand how different premium audio consumers valued different collections of product features and what kind of messaging they would find most appealing. We also needed to locate these consumers in our media targeting to enable segment-based messaging, so we factored those needs into the study design from this early stage.

Obtaining Stakeholder Buy-In

To encourage adoption of the segmentation, we needed to understand what stakeholders wanted *and* what would make them trust the results. While the business was not new to segmentation, some stakeholders found previous models too narrow or difficult to apply to their day-to-day decisions. To design a solution that would meet all of their needs, we expanded on that feedback by asking our stakeholders a series of focused questions:

- **What are your desired outcomes?** What do you hope this segmentation will help you do? How do you envision using it in real decisions?
- **What are your expectations?** How did past models fall short? What would "better" look like this time around? We asked our stakeholders to provide specific examples whenever possible so we could avoid previous pitfalls.
- **What drives your buy-in?** What would make you feel confident that this model actually works for your needs? For non-technical stakeholders, this was especially important: We didn't just need a good model, we needed one that felt trustworthy, usable, and grounded in their world.

Since no two stakeholders had identical needs, it was important to bring everyone together early to define our collective priorities and build shared ownership of the final outcome. This paid off because when our stakeholders saw their needs reflected in the final model, they trusted it, used it, and advocated for its adoption in other areas of the business.

Generating Ideas

As we gathered input from our stakeholders, we quickly learned just how much they wanted to know about our consumers. The initial list included everything from demographics such as age and gender, to habits around consumer audio buying, to feelings about music and other entertainment content, to understanding what other hobbies they have and what other things matter to them.

When trying to build a robust segmentation model, it can be tempting to throw a ton of different variables at it and see what shakes out. However, including too many variables in a latent class model can cause myriad problems including:

- Multicollinearity, which can distort results or produce unstable solutions.
- Poorly differentiated clusters, which can lead to segments that are hard to explain and even harder to act on.

- Narrow or hyper-specific segments that are difficult to find in real-life, especially in outside datasets used for media targeting or gathering additional profiling data.

We conveyed to our stakeholders that, while we understood why all these different elements were important to them, it would not possible to include all of them in the actual segment definitions. For this conversation, we glossed over the technical issues and focused instead on the application issues that could result from including too many variables. These included:

- Harder to explain which consumers landed in each group and why, if the group definitions relied on too many different variables and dimensions
- Harder to explain the differences between each group, if an overabundance of variables led to poorly differentiated classes
- Harder to find the segments "in the wild," when we looked for things like media targeting, if the segment definitions were too narrow or hyper-specific.

Framing the challenges in this very practical and relatable way helped the stakeholders get on board with our goal of limiting the number of variables and dimensions that would form the actual segment definition.

Refine the Dimensions

Once we had the initial "wish list" of variables, we needed to identify which ones should be part of the segment definitions and which should be used to profile the segments after they were identified. To obtain helpful input from our stakeholders, we framed the conversation as an opportunity to clarify:

- The *segmentation variables*, or the dimensions that would *define* the segments and tell us *who we're looking for*.
- The *profiling variables*, or the dimensions that would *further explain* the segments and tell us *who we found*.

As an example, we explained that we knew we would want to learn how to reach younger consumers. But if we included age as a variable in the classification, we would be pushing people together based on their similarity in age, rather than traits that might explain their propensity toward particular products, brands, behaviors, etc. If we used age as a profiling variable instead, we would find groups of consumers with more substantive traits in common and could see if those groups naturally trended younger or older—or if age was less of a factor than we thought.

Drawing upon previous conversations about requirements and expectations, we encouraged the stakeholders to think about it in terms of three questions:

- What characteristics do we want people in each group to **have in common**?
- What characteristics do we want people in different groups to **be different**?
- What characteristics do we want to see **shake out naturally**?

Of course, this is a bit of a simplification because any clustering or classification routine will yield groups that share some characteristics, as well as variation among individuals within a given group. But this was a helpful way to get our stakeholders thinking about how to differentiate between dimensions for classification and dimensions for profiling.

Limiting the dimensions for classification helped to simplify variable selection and was also a strategic decision: There is some well-founded concern that segmentation risks dividing the market into chunks that are very narrow and create target markets that are too small to be useful or to find in the wild. We wanted to be conscientious of that, so limiting the dimensions kept the segments fairly broad, as it resulted in bundles of shared characteristics across different groups while providing useful and interesting nuance. For example, several of our segments shared a love for music, but they were distinguished by a few key ways that love played out in their purchasing behavior that was relevant for the applications we had in mind for the segmentation.

Designing the Survey

Once we selected the dimensions we wanted to reflect in the segmentation, we started thinking about the actual variables, which in our case meant the literal survey questions. While some like demographics were pretty straightforward, others required more nuance, particularly for the psychographic questions used to understand consumers' attitudes.

In this step, it was important to recognize the difference between something like "I shop for the lowest price" versus "I like to get a good value." For example, a nuance we discovered in prior research showed that consumers who shop for a "good value" are not always looking for the lowest price—they are often willing to spend good money on high quality products. To address this, we included several questions to address each of our key dimensions, but which were framed in subtly different ways. We expected these responses to be highly correlated, so we knew we would only use one or two on each dimension, but we wanted to have some options to choose from when building the final model.

We also looked to commercially available data for question framing and language. We knew from our needs and expectations conversations that we would want to find these consumer groups in larger profiling data sets and on media targeting platforms, so the format of those data sets informed our selections for the data collection survey.

When designing the survey, we wanted to incorporate a variety of question types so we would have options when choosing variables for our model. We expected that different question types would produce slightly different results even when measuring similar characteristics, so we used a mix that included:

- **Binary scale questions** to understand consumers' psychographic profile, for example their attitudes on sound quality, brand, etc.
- **MaxDiff** to understand consumers' priorities when making trade-offs on certain things, for example if low price would take precedent over anything else.
- **Single select survey questions** to get more details on particular traits, such as understanding what "easy to use" means to a particular respondent.
- **Multi select survey questions** to understand, for example, what other leisure activities respondents spent their money on.

Reviewing the Variables (Again)

Once the data was collected and cleaned, we ran some basic analyses on the variables. Since we'd included multiple questions that measured similar dimensions, we made another pass at narrowing down potential variables for the model using a few techniques:

- **Correlation analysis:** We anticipated correlation between some variables, but others caught us by surprise. We reviewed the data and chose our variables carefully to avoid multicollinearity whenever possible.
- **Response spread:** Even within a single dimension, the responses for some questions clumped together, while other questions produced a wider variety of answers. We aimed to choose variables that provided a wider spread of distinctive responses to encourage better separation between classes.
- **Language and meaning:** After reviewing the math, we also considered any subtle differences in language or meaning for the questions themselves, to ensure we understood the implications of our choices in the model interpretation.

Developing the Classification

By this point, we had narrowed down the variables sufficiently that nearly any model we chose would be viable from a statistical perspective. A discussion of the metrics for variable selection and assessing latent class models is outside the scope of this case study. But truthfully, in our case, I drank a lot of coffee and examined dozens of candidate models by hand to find *The One* that achieved the right balance between fit statistics, solution stability, explainability, and actionability. This brute-force (or let's say . . . *bespoke*?) approach was possible because the work we put in upfront to streamline the variables limited the search space.

While mathematical rigor was non-negotiable, explainability was equally critical. Being able to explain why people fell into each segment and why it mattered to us as a business was critical for stakeholder buy-in and ensuring that we could apply the segments in "fuzzy" situations where precise classification might not be possible.

Actions and Applications

Once the latent class analysis was complete and we had our segments, we started talking through how to implement them. There were myriad business discussions on topics such as which segments to target, and some of the ways that we put our segments into action included:

- **Building a typing tool for segmenting future survey respondents.** Built using discriminant analysis, this allows us to segment future survey respondents and grow our understanding of feature preferences, price sensitivity, messaging appeal, etc., for each of the segments. Since the underlying classification model relied on a limited number of dimensions, this approach was quite precise.
- **Mapping segments to commercial databases.** Using a decision tree approach, we mapped the segments onto a large commercial database, which allowed us to dig deeper into the behaviors and preferences of each segment without running our own extensive surveys. While less precise than the typing tool due to limitations in the database's search

and filter functions, the questions we'd chosen intentionally mapped well to their responses and our validation study confirmed that we'd done an effective job of finding our segments.

- **Targeting consumers for media buying.** This was the least precise of the options due to technical limitations of the media platforms, but we nonetheless attained good results because the model was highly explainable and we had a great deal of reliable profiling data to draw upon to approximate our segments.

Keeping the dimensions limited and the segments broad made it easier to find them in the wild, which in turn made it easy to apply the new segmentation in areas of the business that could not be served as effectively with previous models—particularly applications like media buying where we could not use a typing tool and needed to rely on less precise methods to find our segments due to the nature of those data sets.

Conclusion

There are many different ways to select variables for segmentation models, and we found that strategic considerations were equally important as the mathematical ones. By gaining clarity on how the segmentation would be used, we could make more intentional decisions throughout the process. Bringing key stakeholders into the conversation from the start helped us build a model that met our business needs and resonated strongly with the people using it.

That early alignment also allowed us to clearly distinguish between the dimensions we'd use to build the segments and the profiling information we'd collect to describe them. As a result, we were able to limit the number and types of variables in the survey, keeping things focused and actionable.

Because we considered stakeholder applications from the outset, we ensured that our segment definitions were specific enough to be meaningful but broad enough to be findable across different types of data sets. With that foundation in place, we could focus on selecting a solution that was both mathematically sound and easily explainable on the dimensions that mattered most.

Tracey Di Lascio- Martinuk

EVALUATING MOBILE-FRIENDLY CONJOINT DESIGNS

PAUL RICHARD MCCULLOUGH
MACRO CONSULTING, INC
DAN YARDLEY
SAWTOOTH SOFTWARE

ABSTRACT

This study compares three conjoint methodologies—Traditional CBC, single-alternative CBC, and a variation called Slider Conjoint (numeric response with a slider interface) across desktop and mobile environments. Using a balanced experimental design and consistent holdouts, we evaluate each method on user engagement, dropout rates, and model quality. Results show trade-offs between respondent enjoyment and model robustness, with Traditional CBC in carousel format showing stronger performance on all model metrics but single-alternative CBC being quicker and more enjoyable for the respondent.

The findings indicate that Traditional CBC with Carousel format achieved a hit rate of 66.8% and an out-of-sample MAE of 5.6%, outperforming the other formats on key model performance metrics. Although the single-alternative formats demonstrated greater respondent enjoyment, their model robustness lagged Traditional CBC. These results highlight the trade-off between data quality and user experience in mobile-first survey environments, providing critical insights for researchers designing mobile-friendly conjoint studies.

INTRODUCTION

The widespread adoption of smartphones has led to a growing proportion of survey respondents completing studies on mobile devices. Industry estimates suggest that between two-thirds and three-quarters of online surveys in the U.S. are now conducted on smartphones. This trend presents challenges for the administration of conjoint studies, which require structured presentation of alternatives that can strain limited screen space.

The evolution of digital platforms and increasing reliance on mobile devices for survey participation have pushed researchers to adapt complex methodologies like conjoint analysis to smaller screens. While traditional CBC methods were designed for desktop environments, user expectations of speed and simplicity have necessitated exploration of alternative approaches. Studies by Smith et al. (2022) and Lee and Park (2021) suggest that mobile-first survey designs not only improve participation rates but also influence data quality, highlighting the need for comparative evaluations like the one undertaken in this paper.

Previous research by Dotson et al. (2023)[1] proposed a simplified single-alternative format for CBC, dubbed "Tinder CBC," as a potentially more mobile-friendly alternative. Their findings indicated that Tinder CBC was quicker, more enjoyable, and yielded similar utilities to

---

[1] See References for full citation.

Traditional CBC. Motivated by their work, this paper evaluates three formats for conducting conjoint studies on mobile and desktop platforms.

While prior work has suggested single-alternative formats may offer a smoother mobile experience, limited empirical comparisons across formats and platforms remain. This study aims to fill that gap.

For the purposes of this paper, the three conjoint formats are defined as follows:

- **Traditional CBC** uses the standard discrete choice format of displaying multiple product profiles simultaneously and asking respondents to choose the one that they would most like to buy, if any. In the Dotson et al. paper, Sawtooth's Carousel format performed marginally better than two other traditional CBC formats. For that reason, the Traditional CBC format we used for mobile devices was Sawtooth's Carousel format.
- **Tinder CBC** format consists of one alternative, rather than multiple alternatives simultaneously. Respondents were asked to swipe left or right on a slider bar below the displayed alternative to show that they would or would not choose to buy that alternative. After sliding the bar left or right, the survey auto advanced to the next screen.
- **Slider Conjoint** is displayed in virtually the identical format as Tinder CBC. However, with Slider Conjoint, the respondent can choose any point on the bar, not just left or right, making Slider Conjoint mathematically equivalent to ratings-based conjoint. Unlike Tinder CBC, after the respondent has slid the tracker ball on the slider scale to the desired location, the respondent must click on the submit button to advance to the next question.

Research Objectives

This study seeks to compare the following three conjoint formats:

- Tinder CBC—Single-alternative choice (swipe left/right)
- Slider Conjoint—Numeric response with a slider interface
- Traditional CBC—Three-alternative plus None, Carousel format on mobile

The goal is to determine which format offers the best trade-off between respondent experience and model performance in mobile environments.

A secondary objective is to examine the relationship between data quality, respondent self-assessed enjoyment of the survey and model performance.

Methodology

Sample was donated for this project by Rep Data[2], a sample provider company.

Invitations to participate were sent by email to potential respondents. Included in the email was a link to the survey. Response rate (proportion of clicks that qualified) was 67% and Conversion Rate (completes/clicks) was 51%. Interview length was 7 minutes. The "bad data"

[2] https://repdata.com/

cell included respondents identified by Rep Data's Research Defender tool as potential sources of low-quality data, providing a necessary contrast to evaluate model robustness under real-world conditions.

Respondents were qualified by the following:

- Residents of USA
- 18–99 years of age
- Pass a standard security question
- Own a smart phone
- Claimed device for completing survey confirmed by checking screen size

Study design consisted of seven cells:

1. Tinder CBC—PC
2. Tinder CBC—Mobile
3. Slider Conjoint—PC
4. Slider Conjoint—Mobile
5. CBC—PC
6. CBC—Mobile
7. CBC- PC—"Bad data"

The first six cells were processed through Rep Data's Research Defender[3] product which cleans the set of potential respondents prior to accessing the survey. The seventh cell consisted of respondents who failed to pass Defender's screening process.

To create conjoint models that would be comparable to each other, we adjusted number of tasks and sample size so that the standard errors of the parameter estimates using random data were roughly equal. This resulted in different sample sizes for each of the three conjoint formats but allowed for the comparisons of conjoint model performance across cells.

Please note that the standard errors from the Logit models are scaled differently than the standard errors from the OLS models, which makes them incomparable directly. The standard errors of the Traditional CBC and Tinder CBC were matched using an aggregate logit model. Then the standard errors of Tinder CBC and Slider Conjoint were matched using an aggregate linear model.

Total sample size was 1,500. The results of the standard error balancing are below in Table 1:

[3] See References for information on Rep Data and Research Defender.

Table 1: Standard Errors

	Traditional CBC	Tinder CBC	Tinder CBC	Slider Conjoint
Sample size	100	400	400	200
Versions	34	34	34	34
Tasks	12	20	20	20
Alternatives/task	3	1	1	1
Design Method	Complete Enumeration (minimal level overlap)			
Test Model	Logit	Logit	OLS	OLS
Std Errors 3 levels[1]	0.044	0.044	0.013	0.018
Std Errors 5 levels[1]	0.068	0.063	0.014	0.02

Tinder and Slider formats were visually consistent across platforms. Traditional CBC used a carousel interface on mobile. Screen Shots of each conjoint format on each platform are below in Figures 1–3:

Figure 1: Tinder CBC—PC and Smart Phone

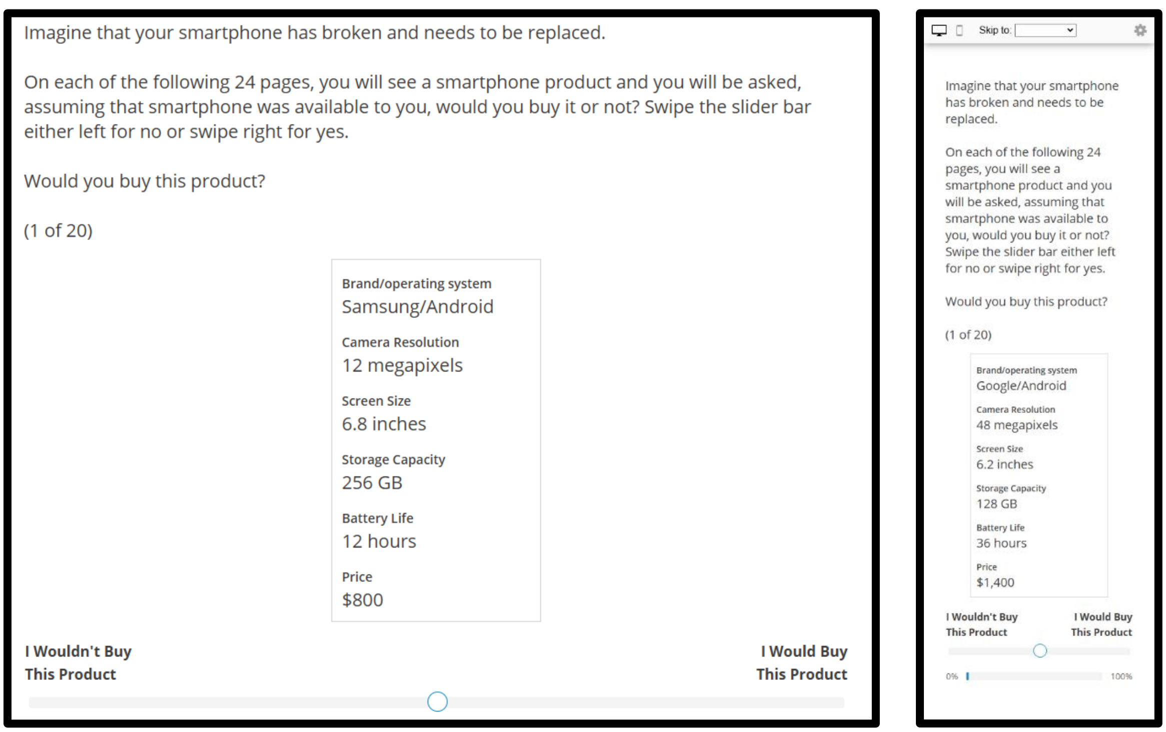


Figure 2: Slider CBC—PC and Smart Phone

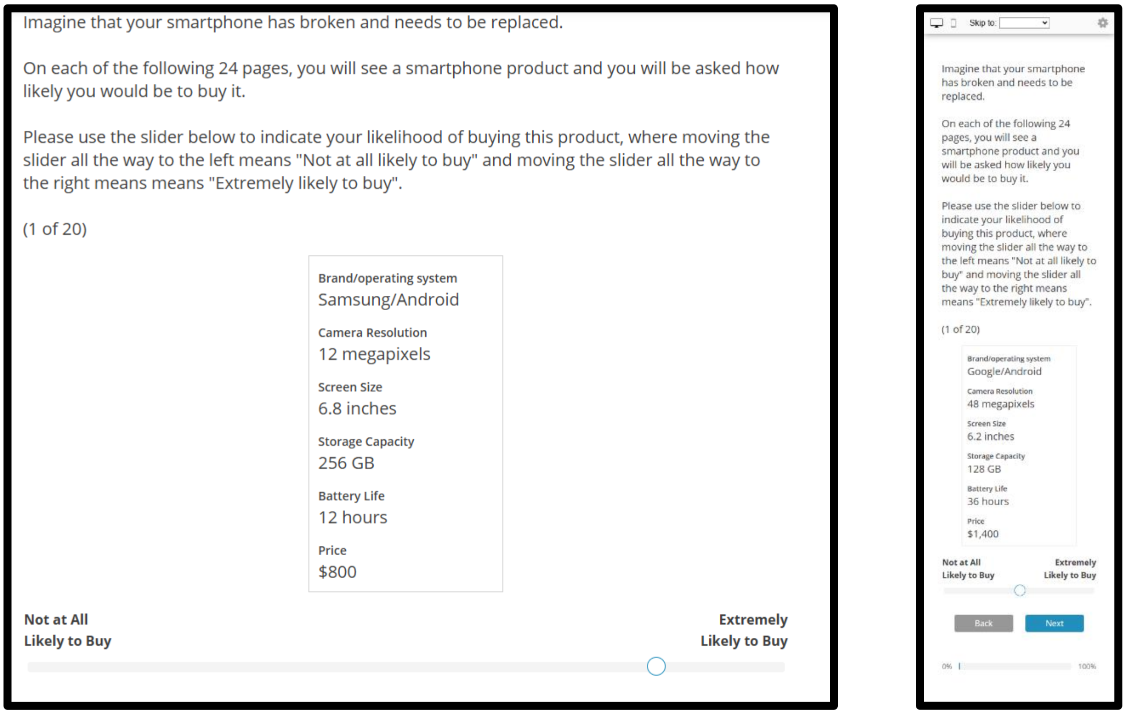
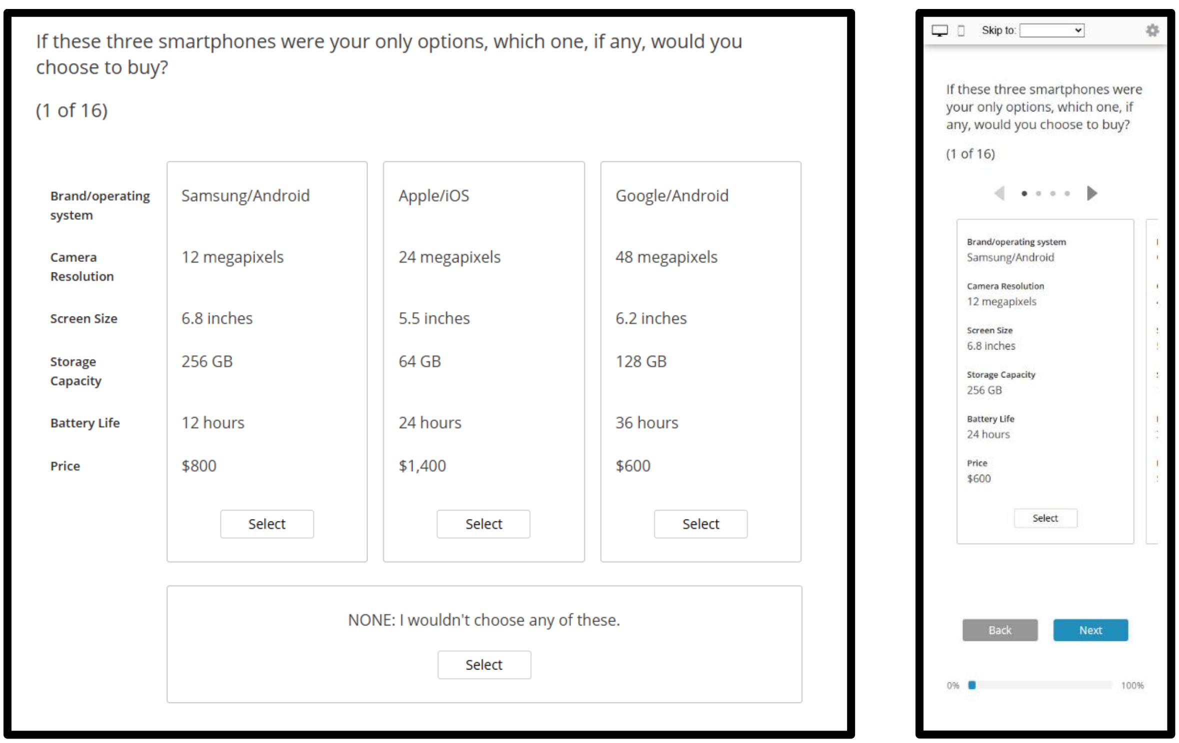


Figure 3: Traditional CBC—PC and Smart Phone



Conjoint Design

The conjoint design consisted of:

- Six attributes
- Three levels per each attribute but price (5 levels)
- No prohibitions
- Four holdout cards, CBC format
- Holdout cards were placed at the end of the Tinder and Slider exercises
- Holdout cards were interspersed among the 12 CBC choice tasks

Attributes and Levels are listed below in Table 2:

Table 2: Attributes and Levels

Brand/operating system	Screen Size	Camera Resolution	Storage Capacity	Charging Speed (to 50%)	Price
Apple/iOS	5.5 inches	12 megapixels	64 GB	30 minutes	$600
Samsung/Android	6.2 inches	24 megapixels	128 GB	60 minutes	$800
Google/Android	6.8 inches	48 megapixels	256 GB	90 minutes	$1,000
					$1,200
					$1,400

Analytic Results

The primary focus of this paper was to measure and compare conjoint model performance across the 6 main cells. To do so, several metrics were used:

- Hit Rates: In each cell, raw choice data from the four Holdout Cards are compared to predicted preference shares to calculate hit rates.
- Out-of-Sample (OOS) MAE: The utilities of one cell are used to estimate Holdout Card preference shares of the other five cells.
- %Fail RLH Check is defined as the percentage of respondents whose RLH is not larger than the RLH expected from random data.
- McFadden's ρ^2: RLH is a goodness-of-fit measure affected by the number of alternatives per task. McFadden's ρ^2 corrects for number of alternatives, making it more appropriate for comparisons across conjoint methods with different numbers of alternatives.
- Attribute Non-Attendance (ANA) refers to a phenomenon in decision-making or choice modeling where individuals or respondents ignore or do not consider certain attributes of

a product, service, or alternative when making a choice, even if those attributes are theoretically relevant. We used the Hess and Hensher (2010) method, as modified by Espinosa-Goded et al. (2021) to identify ANA for each attribute for every respondent.

Using the above measures, it is clear that Traditional CBC using Carousel format on a smart phone performs better than either Tinder CBC or Slider Conjoint on a smart phone:

Table 3: Model Performance

	Mobile		
	Tinder	Slider	CBC
Hit Rate	56.9%	57.9%	66.8%
MAE (OOS)	5.9%	6.5%	5.6%
%Fail RLH Check	21%	11%	7%
McFadden's ρ^2	0.63	0.71	0.77
Attribute Non-Attendance	3.0	2.7	1.9

The higher hit rates and lower dropout rates observed for the Carousel CBC design may stem from the richer choice context, which likely promotes deeper respondent engagement and more considered trade-offs between attributes. However, the marginally higher enjoyment scores for Tinder CBC and Slider Conjoint highlight the trade-off practitioners must consider when prioritizing data quality versus respondent engagement. The significantly lower ANA observed for Carousel CBC indicates that respondents in this format engaged more deeply with the full range of product attributes.

While completion rates are similar, Traditional CBC using Carousel format has a directionally lower dropout rate during the conjoint than either Tinder CBC or Slider Conjoint:

Table 4: Dropout Rates

	Mobile		
	Tinder	Slider	CBC
Completion Rate	89.9%	85.5%	90.2%
Dropout Rate During Conjoint	9.0%	10.5%	5.0%

Additionally, Traditional CBC using Carousel format has lower Attribute Non-Attendance than the other two formats (recall there are six attributes in the conjoint exercise):

Table 5: Attribute Non-Attendance

Mean # of Unattended Attributes	Tinder	Slider	CBC
PC	3.2	2.8	1.9
Mobile	3.0	2.7	1.9

The single alternative formats did perform better regarding interview length and survey enjoyability:

Table 6: Interview Length and Enjoyability

	Mobile		
	Tinder	Slider	CBC
LOI (in minutes)	5.9	6.3	6.7
Enjoyable	79%	79%	68%

Because the Tinder CBC approach offers many fewer alternatives per choice task, i.e., one alternative plus a no buy option, than traditional CBC, there was some interest in comparing flatliners and almost flatliners from Tinder CBC to Traditional CBC.

Although the number of tasks, 20 for Tinder CBC versus 12 for Traditional CBC, might make flatlining less likely for Tinder CBC, the number of alternatives, 1 for Tinder CBC and 3 for Traditional CBC, might make flatlining more likely for Tinder CBC. The data suggest that Tinder CBC respondents are more likely to flatline or almost flatline than Traditional CBC respondents.

Table: Flatliners

	Tinder CBC	Traditional CBC
% Flatliners: None Option	4%	0%
% Flatliners: Alternative	4%	0%

Further, a full 26% of Tinder CBC respondents swiped either left or right at least 85% of the time, suggesting a substantial portion of the Tinder CBC respondents were "almost" flatliners. On the other hand, one Traditional CBC respondent "almost" flatlined 10 of 12 choice tasks by choosing alternative 1 and one "almost" flatlined by choosing None on 10 of 12 choice tasks.

Although the numbers are not large, these data do indicate more of a tendency to flatline with Tinder CBC than Traditional CBC. This can be problematic for utility estimation for Tinder CBC if a respondent swipes either left or right on nearly all the alternatives, as we would learn relatively little about relative preferences for the attribute levels.

Data Quality, Enjoyability and Model Performance

As a secondary objective, we were interested in looking at the relationship between data quality, respondent's perception of survey enjoyability and conjoint model performance. In the below table, we show the results of several Structural Equation Models that were run to determine the relationship between data quality, enjoyability and model performance. McFadden's ρ^2 was the dependent variable and two disaggregate data quality variables that were generated by Rep Data's Research Defender and five self-assessed survey experience variables were used as predictor variables.

The two data quality variables were Threat Assessment and Hyperactivity:

- The **Threat Assessment** variable score reflects the likelihood that a respondent is providing inaccurate or unreliable answers. This score considers factors such as inconsistent or illogical responses, speed of answering, or patterns that suggest a lack of genuine engagement with the survey. A higher Threat Assessment score indicates a higher risk of poor data quality from that respondent.

- The **Hyperactivity** variable score measures how quickly or erratically a respondent is completing the survey, potentially indicating hasty or non-serious participation. It tracks abnormal answering speed, such as rushing through the survey without considering the questions. A high Hyperactivity score suggests that the respondent might be answering too quickly, which could lead to unreliable or poor-quality data.

The survey experience questions were:

- Easy to understand the questions
- Survey length
- Enjoyable
- Interesting
- Willing to participate in another survey

The below table shows the results of the various SEMs:

Table 8: Model Performance Drivers

SEM Std β	Total Sample	PC Platform	Three CBC Cells
Activity	ns	0.054	0.15
Threat	0.033	0.089	0.098
Easy to Understand	0.096	0.101	0.169
Survey length	ns	ns	ns
Enjoyable	ns	ns	ns
Interesting	ns	ns	ns
Willing to participate in another survey	0.041	ns	ns

Interestingly, "Easy to Understand" consistently emerged as a key driver of model performance across platforms and formats, underlining the importance of intuitive survey design for robust utility estimation. Conversely, variables like "Enjoyable" and "Interesting" did not significantly impact model performance, suggesting that respondent enjoyment alone is insufficient to ensure high-quality data.

Please note that the Rep Data Research Defender product so effectively cleaned the data prior to the respondent touching the survey that finding variance in the data quality variables proved somewhat difficult. It was necessary to insert the "bad data" of cell 7 into all models to have sufficient variance for estimation of parameters.

Conclusions

- Mobile CBC Carousel is superior to Mobile Tinder, as well as Slider, with respect to model performance, dropout rates, ANA and required sample size.
- How much a respondent enjoys the survey appears to have no effect on model performance.

- Data quality does have a significant effect on model performance.
- There is a sample penalty for both single alternative approaches compared to traditional CBC.

This study demonstrates the clear advantages of the Carousel CBC format for mobile survey environments. It also raises questions about the scalability of these findings across other product categories or more complex choice tasks. Future research should explore hybrid designs that blend the simplicity of single-alternative approaches with the robustness of traditional CBC, especially as mobile survey technology continues to evolve. Further examination of adaptive designs or real-time data quality checks could also enhance the balance between engagement and model fidelity.

IMPLICATIONS

- Sawtooth's Carousel format for CBC studies does provide better quality choice models than either Tinder CBC or Slider Conjoint. It requires less sample and is only modestly longer to complete.
- But both Tinder and Slider models are still very good.
- If your main priority is model performance, Carousel CBC should be your choice.
- If respondent experience is important, e.g., when interviewing current customers that you don't want to annoy, Tinder CBC will provide a good model and respondents will enjoy the experience more.
- Slider Conjoint and Tinder CBC are virtually tied but Slider Conjoint is a slightly more complicated question, both for the respondent and the analyst, making Tinder CBC preferred between the two.

For practitioners conducting choice modeling in mobile-dominant populations, these findings offer clear guidance. The Carousel CBC format should be the default for high-stakes decisions where model accuracy is paramount, such as pricing optimization or product portfolio management. However, for exploratory research or customer feedback studies where minimizing respondent burden is critical, Tinder CBC offers a viable alternative with acceptable model performance. Balancing these considerations based on research goals and target populations will be essential for maximizing the value of conjoint studies in mobile environments.

ACKNOWLEDGMENTS

The authors thank Rep Data for generously supporting the study with sample and fraud-screening capabilities, Bryan Orme for providing valuable suggestions and the Sawtooth Research Conference for providing a platform to present these findings.

Paul Richard McCullough Dan Yardley

References

Dotson, Jeffrey P., Howell, John, Dotson, Mark, and Lutz, Craig, (2023) "Swipe Right on Simplicity: Examining the Theoretical and Practical Viability of Choice Sets of Size one," 2023 A&I Summit Proceedings, Barcelona, Spain

Espinosa-Goded, M., M. Rodriguez-Entrena and M. Salazar-Ordonez (2021) "A straightforward diagnostic tool to identify attribute non-attendance in discrete choice experiments," Economic Analysis and Policy, 71: 211–226

Hess, S. and D.A. Hensher (2010) "Using conditioning on observed choices to retrieve individual-specific attribute processing strategies," Transportation Research, Part B, 44: 781–790.

Lee, J. & Park, H. (2021). "Respondent Engagement and Data Quality in Mobile Versus Desktop Surveys," *International Journal of Market Research*, 63(2), 185–204.

McFadden, D. (1974) "Conditional logit analysis of qualitative choice behavior." Pp. 105–142 in P. Zarembka (ed.), Frontiers in Econometrics. Academic Press.

Smith, A., Johnson, R., & Kim, S. (2022). "Mobile-Friendly Survey Design: Challenges and Opportunities for Online Research," *Journal of Survey Methodology*, 18(3), 145–161.

Snell, S. (2024) "A Multi-Pronged Approach to Ensure Data Quality," Rep Data White Paper.

COMMENT ON MCCULLOUGH AND YARDLEY

BRYAN ORME
SAWTOOTH SOFTWARE

MANY FLAVORS OF CONJOINT, USUALLY SIMILAR RESULTS

McCullough and Yardley have done a fine job comparing two forms of single-concept CBC to traditional CBC. As many authors have reported at previous Sawtooth conferences, conjoint methods are fairly robust to different forms of presenting attribute information and eliciting responses. Coverage of different approaches have included:

- Card-sort, full-profile ratings-based and ranking-based conjoint
- Partial-profile adaptive conjoint (ACA)
- Full-profile CBC (discrete choice, best-worst, or chip allocation)
- Partial-profile CBC
- Adaptive CBC (ACBC)
- Tournament-style CBC

The context and attention/consideration elicited from respondents differs from method to method. The results (part-worth utilities and predictions from market simulators) are usually quite similar, though not identical. (As one example, ACA tends to understate the importance of price). Different conjoint methods vary in terms of respondent friendliness, statistical efficiency, response error (scale), and predictive validity.

In general, we recommend conjoint methods that are respondent friendly, statistically efficient, and mimic the actual purchase process as closely as possible, with the aim of predictive validity. There are of course tradeoffs involved when balancing these goals.

"TINDER"-STYLE CBC

In terms of respondent friendliness, statistical precision, and task realism, Tinder-style conjoint does well in terms of respondent friendliness (McCullough and Yardley 2025, Dotson et al. 2023), poorly on statistical precision, and probably poorly as well in terms of mimicking most purchase processes.

HOW NONE USAGE AFFECTS STATISTICAL PRECISION

We can use Sawtooth's Test Design procedure to measure the amount of information (and the statistical precision that follows) for CBC experiments. The test procedure simulates a population of random-answering respondents, estimates the part-worth utilities using aggregate logit, and using the variance-covariance matrix that follows computes a summary relative efficiency metric (relative D-efficiency).

The percent of None choices can have a large effect on the precision of the part-worth utilities for the attributes and levels of interest. If respondents choose only the None alternative, then we learn very little about their preferences for the attributes and levels of interest.

A safety net for CBC researchers is the dual-response None approach, where respondents are first forced to choose among a set of product concepts (defined on the attributes and levels of interest), then asked whether they would in reality buy the concept they chose as best. No matter the None percentage, the preliminary forced choice among product alternatives ensures that we obtain robust tradeoff information about the attributes and levels of interest.

With that introduction and background, let's now consider Tinder-style conjoint.

Statistically Inferior Experimental Design, Risky Data Collection

Tinder-style conjoint is statistically inferior and risky compared to standard CBC. I use the word risky, because depending on the tendency toward using None for a given respondent, you can end up with very little information about the attributes and levels of interest.

But isn't this no different from standard CBC with traditional None, you might object? If respondents mainly answer with the None choice in traditional CBC, you also don't learn much about the attributes and levels of interest. However, there is double the risk with Tinder-style CBC. Not only do we have to worry about respondents swiping too many cards to the *left* (indicating None), but we're also sunk if respondents swipe too many cards to the *right* (indicating that they like nearly all of them).

In the case of single concept versus the None, statistical information regarding the attributes and levels of interest is maximized when respondents choose the None at a rate of 50%. As the None approaches either 0% or 100% we obtain very little information about the attribute levels of interest. Think about it: if a respondent swipes all 20 concepts to the right, what have we learned about their tradeoffs among the attributes and levels of interest? *Nothing.* Same goes if they swipe all 20 concepts to the left.

With that in mind, we can simulate the relative statistical precision (relative D-efficiency) for Tinder-style CBC compared to standard CBC (showing three concepts) with dual-response None. We can do so under different assumptions of the degree to which the Tinder-style respondents swipe either mostly left or mostly right.

Relative D-Efficiency by Percent of Left or Right Swipes

Tinder CBC Percent Choices Left (or Right)	*Relative Efficiency (Task Equalized)	*Relative Efficiency (Time Equalized)
95%	5.5%	10.9%
90%	10.3%	20.7%
85%	14.6%	29.3%
80%	18.4%	36.7%
75%	21.5%	43.0%
70%	24.1%	48.2%
65%	26.1%	52.2%
60%	27.5%	55.1%
55%	28.4%	56.8%
50%	28.7%	57.4%

*Relative to dual-response CBC with 3 concepts/task

Under the most favorable conditions where Tinder-style respondents swipe left 50% of the time, the task-equalized experiment is 28.7% as efficient as CBC with dual-response None. When time-equalized (given that respondents can complete Tinder-style tasks about twice as fast as CBC tasks), the best-case relative efficiency is 57.4%.

McCullough and Yardley found that 26% of their respondents swiped the same direction 85% or more. Referring to the table above, that means 26% of their sample is at the 29% or lower relative efficiency level.

Concluding Thoughts

Conjoint methods tend to be fairly robust to different context and response style formats. This newcomer, Tinder-style conjoint, also leads to fairly similar results as standard CBC; though I worry about response simplification—McCullough and Yardley showed that respondents pay attention to fewer attributes with Tinder-style conjoint. Perhaps Tinder-style conjoint better reflects screening and consideration rather than final choice?

Tinder-style conjoint is also a riskier approach than traditional CBC with dual-response None, as respondents who tend to answer mainly with left or right swipes lead to increasingly imprecise estimates of preferences for the attributes and levels of interest.

Bryan Orme

TOKEN-BASED CONJOINT—A NEW FRAMEWORK FOR TOO MANY ATTRIBUTES

MEGAN PEITZ
TREVOR OLSEN
NUMERIOUS INC.

ABSTRACT

This paper introduces Token-Based Conjoint (TBC), a novel approach for conducting conjoint analysis when the number of product features is exceptionally large (30+ attributes). TBC draws inspiration from MaxDiff and choice-based conjoint (CBC), asking respondents to evaluate small subsets of features by allocating tokens to their top preferences, followed by a purchase likelihood question. This structure enables respondents to meaningfully engage with complex feature sets while generating data amenable to additive modeling.

We validated TBC across three studies involving up to 40 features, assessing its predictive power through both in-sample and out-of-sample tasks. Results demonstrated strong accuracy and offered clear insights into feature-level drivers of uptake. We also explored ways to scale TBC further, testing simplified task structures and a weighted modeling approach to mitigate overfitting when task exposure varies.

TBC is not a one-size-fits-all method. In scenarios where price is a central focus, such as revenue optimization or willingness to pay per feature, other methods may be more suitable.

Overall, TBC provides a practical, respondent-friendly solution for high-dimensional feature prioritization. It is especially well-suited for uncovering additive value when pricing isn't the primary objective—particularly in settings where diminishing returns and utility saturation are key considerations.

INTRODUCTION

One of the most common questions in product research is: *"Which combination of benefits will make people buy my product?"* This was exactly the challenge posed by a client who wanted to identify the optimal feature bundle for a new offering. Conjoint analysis was the obvious starting point as it's designed to reveal how combinations of features drive preference. But when we asked how many benefits they wanted to test, the answer came back: *"Maybe 30, even up to 40."*

At that scale, traditional conjoint methods start to break down. Presenting respondents with dozens of attributes leads to cognitive overload and unreliable data. Even advanced designs that try to accommodate large attribute sets have issues: they often produce profiles with unrealistic balance (e.g., every product having the same number of features "on") and fail to capture diminishing returns when stacking multiple high-utility features.

MaxDiff, or best-worst scaling, is another common alternative. It handles long feature lists well and can rank features by importance. But MaxDiff results are not additive, and do not model purchase likelihood. It doesn't tell us how feature *combinations* perform. We might learn that

Feature A is more appealing than Feature B, but we can't predict how much more attractive a product becomes if it includes *both*. Without additive structure, we can't simulate bundle-level uptake.

This gap between conjoint and MaxDiff led us to develop Token-Based Conjoint (TBC)—a new framework that blends the strengths of both methods. In a TBC, respondents evaluate small subsets of features, "spend tokens" on their favorites, and then indicate whether that set would be enough to trigger a purchase. Repeating this across multiple feature subsets allows us to gather additive data on feature value without overwhelming respondents.

In this paper, we introduce the TBC approach and share findings from three large-scale studies involving up to 40 features. We begin by detailing how TBC works, including task design, implementation, and how data is coded for modeling. We then evaluate its performance using fixed holdout tasks and out-of-sample prediction, demonstrating strong predictive accuracy. Next, we explore how TBC can be scaled for even larger or more complex scenarios by modifying the task structure and applying a likelihood weighting strategy to maintain balance. We conclude with key takeaways, limitations, and practical guidance for applying TBC in future product research.

METHODOLOGY: THE TOKEN-BASED CONJOINT APPROACH

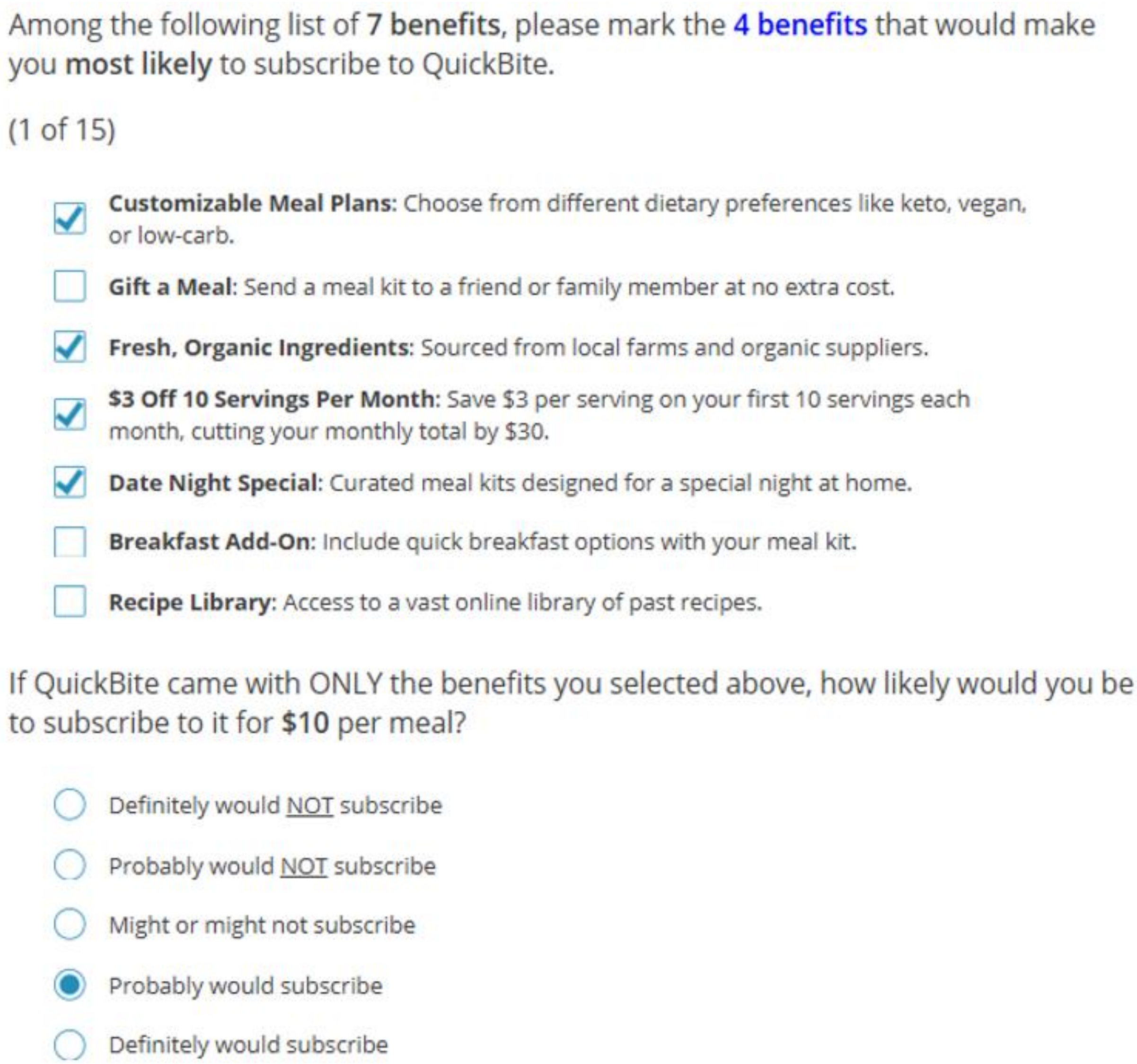
Design of the TBC Task

In a Token-Based Conjoint (TBC) exercise, respondents evaluate manageable subsets of product features and select their top preferences. Each screen presents $\boldsymbol{n}$ features, and respondents are asked to choose the $\boldsymbol{k}$ features they value most. A key innovation is that $\boldsymbol{k}$ is dynamic and varies from task to task. One screen might ask the respondent to pick just one feature ($\boldsymbol{k} = 1$), the next might ask for two ($\boldsymbol{k} = 2$), and so on, up to $\boldsymbol{k} = \boldsymbol{n}\text{-}1$.

This "allocate $\boldsymbol{k}$ tokens out of $\boldsymbol{n}$" design mirrors real-world decision-making: consumers rarely get everything they want and must make trade-offs. By varying $\boldsymbol{k}$ across tasks, we allow respondents to signal both their absolute top choices and the broader set of features they find valuable under different constraints.

After each selection, we include a 5-point purchase likelihood question. This follow-up question is a simplified, purchase-based framing of the None option, similar to dual-response None formats used in conjoint. It captures whether a respondent's selected bundle is strong enough to trigger a real purchase decision. We used a five-point scale ranging from "Definitely would subscribe" to "Definitely would not subscribe," providing a nuanced measure of intent.

Image 1: Example TBC Screen



Among the following list of **7 benefits**, please mark the **4 benefits** that would make you **most likely** to subscribe to QuickBite.

(1 of 15)

- [x] **Customizable Meal Plans:** Choose from different dietary preferences like keto, vegan, or low-carb.
- [ ] **Gift a Meal**: Send a meal kit to a friend or family member at no extra cost.
- [x] **Fresh, Organic Ingredients**: Sourced from local farms and organic suppliers.
- [x] **$3 Off 10 Servings Per Month**: Save $3 per serving on your first 10 servings each month, cutting your monthly total by $30.
- [x] **Date Night Special**: Curated meal kits designed for a special night at home.
- [ ] **Breakfast Add-On**: Include quick breakfast options with your meal kit.
- [ ] **Recipe Library**: Access to a vast online library of past recipes.

If QuickBite came with ONLY the benefits you selected above, how likely would you be to subscribe to it for **$10** per meal?

- ○ Definitely would NOT subscribe
- ○ Probably would NOT subscribe
- ○ Might or might not subscribe
- ● Probably would subscribe
- ○ Definitely would subscribe

Why This Design?

The TBC task is designed to balance cognitive simplicity with realistic trade-offs. By limiting each screen to ***n*** features, we avoid overwhelming respondents with the full list of attributes. No matter how long the master list is, the respondent only evaluates a manageable subset at a time.

The use of a dynamic ***k***, where respondents select different numbers of features on each screen, adds further depth. It allows us to observe preferences for both small, focused bundles and larger, more inclusive ones, all within the same exercise. This variation also prevents over-selection bias: even highly desirable features must compete to be the *one and only* pick when ***k*** *= 1*, but may be chosen alongside others when ***k*** *= 5 or 6*. This helps uncover diminishing returns, where the value of a feature might be high in isolation but less impactful when bundled with similar benefits.

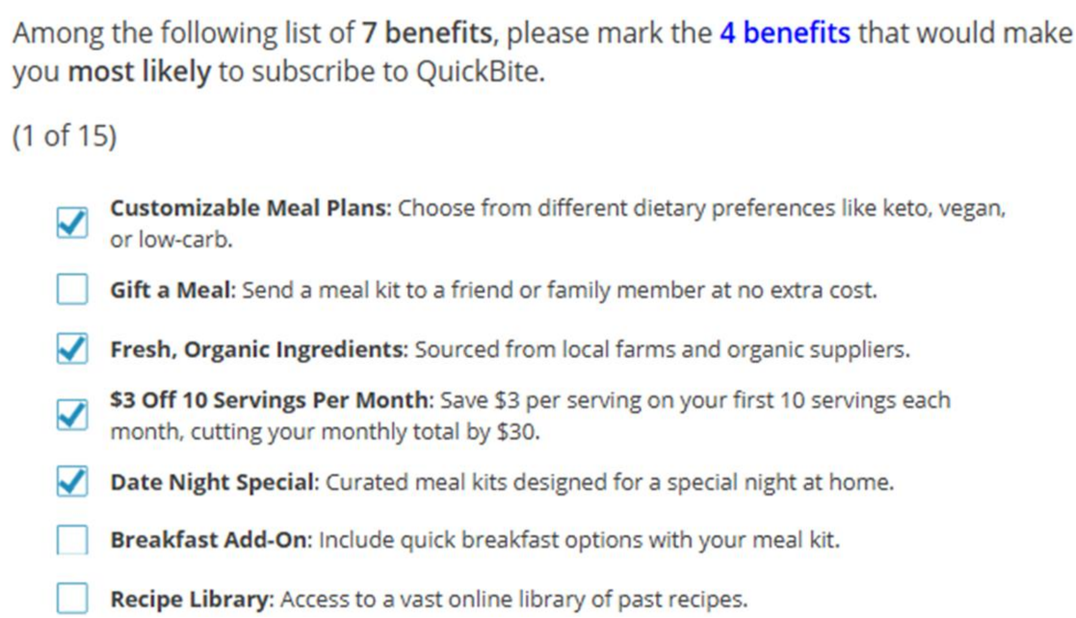
The purchase likelihood follow-up question grounds the task in a real-world decision: *Would this specific bundle be good enough to buy?* Unlike traditional best-worst or rank-based tasks, this step introduces an absolute threshold. Some bundles pass the bar; others don't. That binary framing—*buy or not buy*—feeds additive purchase information into the model, much like the None option in a traditional conjoint, allowing us to simulate both relative feature appeal and overall purchase likelihood.

Implementation

We implemented the TBC exercise using Sawtooth Software's Lighthouse Studio. Technically, we configured the task as a looped free-format question based on a MaxDiff design with a constructed list to manage the dynamic number of selections.

A balanced MaxDiff design was generated with the total number of features, ensuring that each feature appeared across screens a consistent number of times and in varied combinations with others. Instead of asking respondents to select a "best" and "worst" item (as in standard MaxDiff), we modified the prompt to say:

Image 2: Token Exercise



We used Lighthouse's constructed list functionality to pipe in the correct ***k*** value for each screen dynamically.

Once a respondent selected their top ***k*** features, they were immediately shown a follow-up question (See Image 3).

Image 3: Purchase Likelihood Follow-Up Question

This follow-up links the feature selection directly to an outcome, injecting purchase intent into the dataset alongside preference information.

Coding the Data for Analysis

TBC responses can be analyzed using a standard discrete choice modeling framework by treating each screen as a constructed choice task among possible feature bundles. When a respondent sees $\boldsymbol{n}$ features and selects $\boldsymbol{k}$ of them, we interpret their chosen set as a single "alternative" in a choice task.

To define the full choice set, we enumerate all possible combinations of size $\boldsymbol{k}$ that could have been selected from the $\boldsymbol{n}$ features shown. For example, if $\boldsymbol{n} = 5$ and $\boldsymbol{k} = 2$, there are "5 choose 2"

$$\binom{5}{2} = 10$$

possible 2-feature combinations. A respondent chose one of these combinations and implicitly rejected the other nine.

For analysis, we can construct a choice task with these 10 alternatives: assign a 1 to the selected bundle and 0 to the others. This allows us to estimate the model using approaches like multinomial logit (MNL) or hierarchical Bayes MNL, treating each bundle as an observed choice from a well-defined set of alternatives.

Image 4: Coding of a 5 Choose 2 Task
where Respondent Picked Item 2 and 4

Task	Alternative	Item #					Winner	
		1	2	3	4	5		
1	1	0	1	0	1	0	1	Winning Pair
1	2	1	1	0	0	0	0	
1	3	0	1	1	0	0	0	
1	4	0	1	0	0	1	0	
1	5	1	0	0	1	0	0	
1	6	0	0	1	1	0	0	
1	7	0	0	0	1	1	0	
1	8	1	0	1	0	0	0	No winning items
1	9	1	0	0	0	1	0	
1	10	0	0	1	0	1	0	

Incorporating the Purchase Likelihood (None) Response

To capture not just what features respondents prefer, but whether their chosen bundle is actually good enough to purchase, we include a follow-up purchase likelihood question after each selection. This allows us to incorporate a None option into the modeling framework, indicating whether the respondent would buy the bundle or not.

The standard approach, often referred to as dual-response none coding (Diener et al., 2006), handles this as follows:

- If the respondent would not purchase the bundle, we model two tasks:
  1. A choice among the possible feature combinations (with the chosen bundle coded as selected), and
  2. A second task including None as an alternative, coded to show the respondent explicitly chose "None."

- If the respondent would purchase the bundle, we model just one task: a choice that includes the None option, with the selected bundle preferred over it.

This structure helps the model estimate both relative preferences among features and the absolute threshold required to drive a purchase.

In our implementation, we tested an alternative approach called the naïve none. Here, we generate two modeled tasks for every screen, regardless of the respondent's purchase likelihood:

1. A standard choice among feature bundles (as described above), and
2. A binary task comparing the chosen bundle directly against None.
   If the respondent said they would purchase, the bundle is marked as preferred; if not, None is.

Unlike the standard method, the naïve none consistently creates two analytic choice sets for all screens. This uniformity ensures that each respondent contributes the same amount of information—one feature selection task and one purchase decision task per screen.

Image 5: Adding the None Coding

	Task	Alternative	Item #					None	Winner
			1	2	3	4	5		
Respondent says they would buy	2	1	0	0	0	0		1	0
	2	2	0	1	0	1	0	0	1
Respondent says they would NOT buy	2	1	0	0	0	0		1	1
	2	2	0	1	0	1	0	0	0

Handling the 5-Point Purchase Likelihood Scale

To capture nuanced purchase intent, we asked respondents to rate how likely they would be to subscribe or purchase using a five-point scale, ranging from *"Definitely would subscribe"* to "*Definitely would not subscribe*." This opened the door to different ways of translating those responses into binary "buy vs. not buy" decisions in the model.

We explored three coding strategies:

- Top-Box None: Only the most enthusiastic response, *"Definitely would subscribe,"* was treated as a purchase. All other responses were coded as a "no purchase," meaning "None" is preferred in the model.
- Top-2-Box None: This more inclusive threshold counted both *"Definitely"* and *"Probably would subscribe"* as indicating a purchase. Any weaker response was still treated as a "no purchase."
- Calibrated Likelihood: Instead of a binary cutoff, this approach assigns partial credit to purchase likelihood. For example, *"Definitely"* might correspond to a 70% likelihood of buying, and *"Probably"* to 30%. In this case, the model distributes the outcome weight proportionally between the selected bundle and the None alternative.

Image 6: Different None Coding Strategies

The choice you made in the DR	Top Box	Top 2 Box	Calibrated
1 – Definitely would not subscribe	0	0	0
2 – Probably would not subscribe	0	0	0
3 – Might or might not subscribe	0	0	0
4 – Probably would subscribe	0	1	.3
5 – Definitely would subscribe	1	1	.7

Image 6A: Coding the Model with a Top Box Choice

	Task	Alternative	Item #					None	Winner	
			1	2	3	4	5			
Respondent picks top box in the top box model	2	1	0	0	0	0		1	0	*Would NOT be included if Diener Coding*
	2	2	0	1	0	1	0	0	1	
Respondent did not pick top box in top box model	2	1	0	0	0	0		1	1	
	2	2	0	1	0	1	0	0	0	

Image 6B: Coding the Model with a Top 2 Box Choice

	Task	Alternative	Item #					None	Winner	
			1	2	3	4	5			
Respondent picks top 2 box in the top 2 box model	2	1	0	0	0	0		1	0	*Would NOT be included if Diener Coding*
	2	2	0	1	0	1	0	0	1	
Respondent did not pick top 2 box in top 2 box model	2	1	0	0	0	0		1	1	
	2	2	0	1	0	1	0	0	0	

Image 6C: Coding the Model with a 70/30 Calibration

	Task	Alternative	Item #					None	Winner
			1	2	3	4	5		
Respondent picks top box in the calibrated model	2	1	0	0	0	0		1	.3
	2	2	0	1	0	1	0	0	.7
Respondent picks 2nd box in the calibrated model	2	1	0	0	0	0		1	.7
	2	2	0	1	0	1	0	0	.3
Respondent did not pick top 2 box in calibrated model	2	1	0	0	0	0		1	1
	2	2	0	1	0	1	0	0	0

These none-response coding schemes were part of a secondary exploration, aimed at understanding how the model responds to different purchase thresholds. While performance varied slightly across approaches, it's important to note that accuracy alone doesn't reveal which threshold is "correct." Instead, these options provide flexibility for researchers to align their modeling assumptions with prior knowledge or external calibration targets.

Baseline TBC Configuration

Throughout this paper, we refer to the "baseline" TBC model as the original version used across all three studies. It includes:

- The full ***n***-choose-***k*** task structure (i.e., the complete set of possible bundles on each screen)
- Naïve none coding (a binary choice between the selected bundle and None after every task)

STUDY DESIGN AND DATA

Real-World Data

We evaluated the TBC method across three real-world client projects, anonymized and reframed for this paper as if they were conducted for a fictional meal subscription brand called "QuickBite." While all feature descriptions were generalized, the core design structure and data characteristics were preserved.

Each project involved a long list of potential product benefits:

- Projects A and B tested 38 features each
- Project C tested 40 features

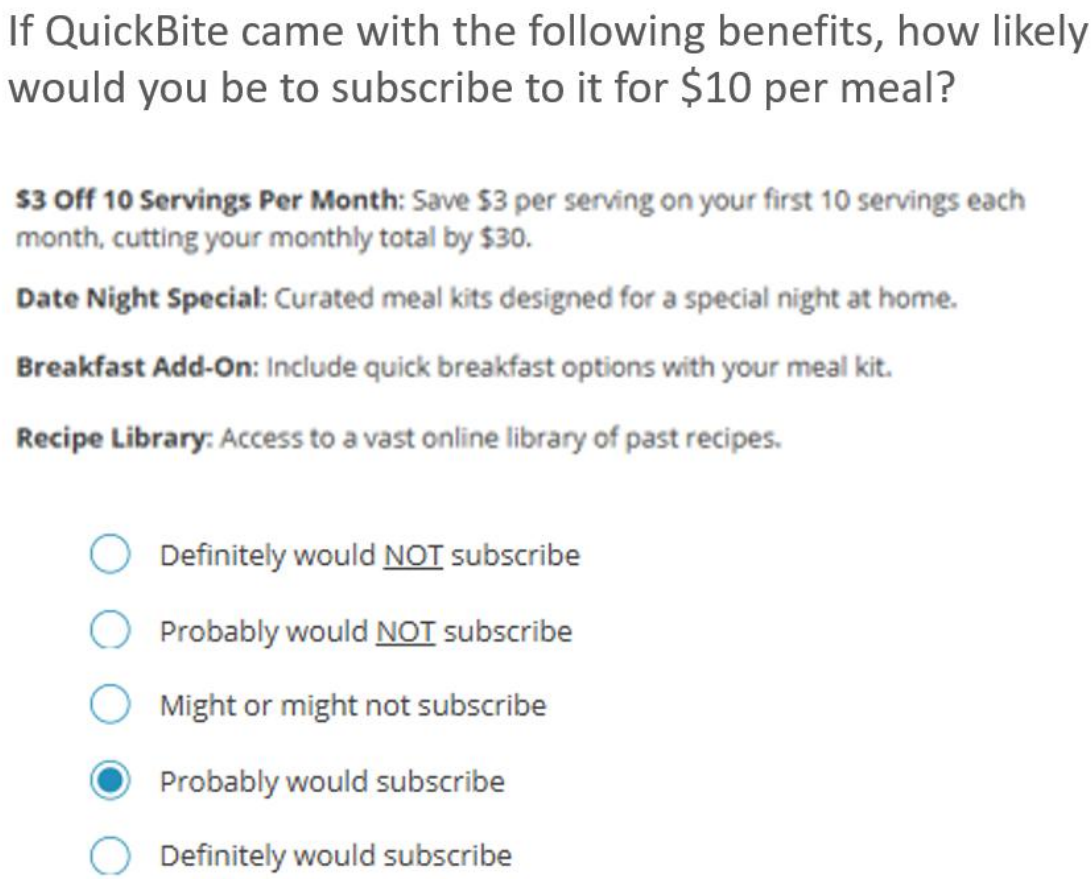
In every case, respondents completed a TBC exercise where $n = 7$ features were shown per screen, and they were asked to select a dynamic number of top features (k) ranging from 1 to 6. For example, one screen might ask for their single top pick ($k = 1$), while another might ask for their top five ($k = 5$).

- Projects A and B included 15 selection tasks per respondent
- Project C included 12 tasks
- Sample sizes:
    - Project A: ~500 (UK sample)
      Project B: ~500 (US sample)
      Project C: 812 (US sample)

Holdout Tasks for Validation

To benchmark predictive accuracy, we included two fixed holdout tasks in each survey. These tasks were the same for all respondents and presented a realistic bundle of features along with a purchase intent question, such as:

Image 7: Sample Holdout Screen



Respondents answered using the same 5-point scale used elsewhere in the survey. Because everyone saw the same holdouts, we can directly compare actual responses to the model's predicted purchase likelihood, making these tasks a useful anchor for validation.

Validation Strategy

We used two key validation methods to evaluate model performance:

1. Holdout Task Prediction (In-Sample and Out-of-Sample)

   Using each respondent's estimated part-worth utilities, we predicted their probability of subscribing to each holdout bundle. We then compared these predictions to actual responses, calculating mean absolute error (MAE) as the metric.

   - *In-sample*: Model estimated on the full dataset
   - *Out-of-sample*: 4-fold cross-validation, where each respondent is predicted using a model trained on a different 75% of the sample

2. Likelihood-Based Validation (Preference Ranking)

   To assess how well the model captures relative preference ordering, we ran a leave-one-task-out test. For each screen number, we held it out for all respondents from estimation. We then used the model to compute the log-likelihood of the actual choice in that task.

Repeating this across all tasks and averaging the results gave us a measure of how well the model ranks bundles—even for unseen combinations.

Together, these two metrics—MAE and log-likelihood—help evaluate both calibration (predicting how likely someone is to buy) and discrimination (identifying which benefits are most preferred).

RESULTS

TBC Baseline Performance and None Coding

We compared two None coding structures: the standard (or "Diener") coding and the alternative naïve none, as previously described. Although both approaches are based on identical survey responses, they differ in how choice data is structured for modeling.

The results were consistent across all three projects: the naïve none approach yielded substantially lower out-of-sample MAEs on the fixed holdout tasks. For example, in Project A, the MAE dropped from 11.90 under standard coding to just 3.33 with naïve none. Projects B and C showed similarly strong improvements.

Image 8: In Sample Mean Absolute Errors

	Dual Response Coding	Naïve None	Diener None
Project A	Top Box	1.45	11.70
	70/30 Calibrated	3.16	16.01
	Top 2 Box	3.28	17.12
Project B	Top Box	4.46	15.07
	70/30 Calibrated	2.57	18.97
	Top 2 Box	1.80	12.76
Project C	Top Box	0.19	4.57
	70/30 Calibrated	0.42	4.38
	Top 2 Box	6.09	7.02

Image 9: Out-of-Sample Mean Absolute Errors

	Dual Response Coding	Naïve None	Diener None
Project A	Top Box	3.33	11.90
	70/30 Calibrated	3.43	16.30
	Top 2 Box	5.80	17.20
Project B	Top Box	5.97	15.00
	70/30 Calibrated	4.32	19.19
	Top 2 Box	5.14	12.77
Project C	Top Box	1.62	4.56
	70/30 Calibrated	1.07	4.64
	Top 2 Box	6.77	7.01

This suggests that the binary structure and consistent task formatting of the naïve none, where every response explicitly informs purchase intent, help the model more precisely estimate individual thresholds, leading to stronger predictive accuracy.

We then assessed whether this coding choice affected the model's ability to rank benefits, a critical function in preference modeling. Using a leave-one-task-out log-likelihood validation, we evaluated how well the model could recover held-out choices across random tasks.

Here, both coding strategies performed similarly. Log-likelihoods were consistent across all three projects, indicating that ranking ability was not compromised by adopting the naïve none approach.

Image 10: In Sample Benefits Ranking Choice Sets:
Sum(Log Likelihoods)/1000
(1st choice set from each screen)

	Dual Response Coding	**Naïve None**	**Diener None**
Project A	Top Box	-1.42	-1.42
	70/30 Calibrated	-1.40	-1.40
	Top 2 Box	-1.42	-1.42
Project B	Top Box	-1.38	-1.38
	70/30 Calibrated	-1.36	-1.38
	Top 2 Box	-1.38	-1.38
Project C	Top Box	-2.25	-2.25
	70/30 Calibrated	-2.23	-2.31
	Top 2 Box	-2.27	-2.27

Image 11: In Sample Naive None Choice Sets:
Sum(Log Likelihoods)/1000
(2nd choice set from each screen)

	Dual Response Coding	**Naïve None**	**Diener None**
Project A	Top Box	-0.20	-0.34
	70/30 Calibrated	-0.35	-0.70
	Top 2 Box	-0.14	-0.26
Project B	Top Box	-0.21	-0.34
	70/30 Calibrated	-0.37	-0.73
	Top 2 Box	-0.13	-0.21
Project C	Top Box	-0.39	-0.62
	70/30 Calibrated	-0.52	-1.01
	Top 2 Box	-0.33	-0.52

The naïve none coding structure improved prediction accuracy on fixed tasks, reflected in lower MAEs, without sacrificing the model's ability to rank alternatives. For that reason, all analyses that follow in this paper use the naïve none approach.

Scaling to "Too Many" Attributes: Option 2 and 3 Experiments

Having validated the baseline TBC approach—referred to here as **Option 1**, which uses the full ***n***-choose-***k*** choice task with naïve none coding—we turned to a forward-looking question: *How can we scale TBC to handle even larger feature sets or bundle sizes?*

While Option 1 worked well for up to 40 features, it may become impractical in scenarios where respondents are asked to choose from or build larger bundles (e.g., 10 out of 15 features), due to the exponential growth in the number of possible combinations (15 choose 10 = 3,003 combinations!). To address this, we explored four alternative task structures aimed at either reducing the number of alternatives shown or restructuring the data to be more digestible for the model.

These task structure variations are summarized below:

- **Option 2A: Append Pairwise Comparison Tasks**—Adds binary comparisons between selected and unselected features.
- **Option 2B: Comparison-Only Tasks**—Drops the original full task and keeps only the binary comparisons.
- **Option 3A: Trimming Unchosen Combinations**—Reduces the number of alternatives by removing bundles made only of unchosen features.
- **Option 3B: Chunking by Chosen Features**—Splits the original choice set into smaller blocks, each centered around a selected feature.

In the following sections, we describe each of these options in more detail and evaluate how well they retain predictive accuracy compared to the baseline.

Option 2A: Append Pairwise Comparison Tasks

This approach augments the original choice data by adding additional tasks that isolate the contribution of each selected feature. After generating the full set of ***n***-choose-***k*** combinations for a given screen, we append ***k*** *new tasks*—one for each chosen feature. In these tasks, each selected feature is directly compared against all features that were *not* selected.

For example, if a respondent chose Features 2 and 4 out of a set of 5, we would add one task comparing Feature 2 to Features 1, 3, and 5, and another comparing Feature 4 to that same set. The intent is to provide the model with clearer, more focused signals: rather than inferring preferences from a single multi-feature choice, we explicitly show that Feature 2 was preferred over 1, 3, and 5—and the same for Feature 4.

Image 12: Option 2A Coding of a 5 Choose 2 Task where Respondent Picked Item 2 and 4

Task	Alternative	Item #					Winner
		1	2	3	4	5	
1	1	0	1	0	1	0	1
1	2	1	1	0	0	0	0
1	3	0	1	1	0	0	0
1	4	0	1	0	0	1	0
1	5	1	0	0	1	0	0
1	6	0	0	1	1	0	0
1	7	0	0	0	1	1	0
1	8	1	0	1	0	0	0
1	9	1	0	0	0	1	0
1	10	0	0	1	0	1	0
2	1	0	1	0	0	0	1
2	2	1	0	0	0	0	0
2	3	0	0	1	0	0	0
2	4	0	0	0	0	1	0
3	1	0	0	0	1	0	1
3	2	1	0	0	0	0	0
3	3	0	0	1	0	0	0
3	4	0	0	0	0	1	0

Winning Pair

Item 2 (winner) versus item 1, 3, and 5 (losers)

Item 4 (winner) versus item 1, 3, and 5 (losers)

Option 2B: Comparison-Only Tasks (Dropping the Full Combination Task)

This variation takes Option 2A a step further by eliminating the original ***n***-choose-***k*** task entirely. Instead of modeling the full set of possible combinations, Option 2B retains *only* the ***k*** pairwise comparison tasks for each screen.

Using the earlier example—where a respondent selects Features 2 and 4 out of 5—this approach drops the 10-alternative combination task and keeps just the two simpler tasks: one comparing Feature 2 to unchosen features (1, 3, 5), and another doing the same for Feature 4.

The benefit is a significant reduction in complexity: fewer alternatives, smaller task sizes, and potentially faster estimation. The trade-off is that we lose the context of the full joint selection. This approach assumes the individual comparisons are sufficient to capture the underlying preference structure.

Image 13: Option 2B Coding of a 5 Choose 2 Task where Respondent Picked Item 2 and 4

Task	Alternative	Item #					Winner
		1	2	3	4	5	
2	1	0	1	0	0	0	1
2	2	1	0	0	0	0	0
2	3	0	0	1	0	0	0
2	4	0	0	0	0	1	0
3	1	0	0	0	1	0	1
3	2	1	0	0	0	0	0
3	3	0	0	1	0	0	0
3	4	0	0	0	0	1	0

Item 2 (winner) versus item 1, 3, and 5 (losers)

Item 4 (winner) versus item 1, 3, and 5 (losers)

Option 3A: Trimming Unchosen Combinations

While Options 2A and 2B expand the data by adding tasks, Option 3A takes a reductive approach: it simplifies the choice set by removing alternatives made up entirely of non-chosen features. This raises a key question: *Does the model need to see combinations that respondents clearly ignored? Or is it enough to focus on alternatives that include at least one selected item to understand preference signals?*

In this approach, we start with the full ***n***-choose-***k*** set, then drop any alternative that contains *only* unchosen items. For example, if a respondent selects Features 2 and 4 from a set of 5, combinations like {1,3}, {1,5}, and {3,5}—which include none of the selected features—are excluded, reducing the total from 10 to 7.

The trade-off is that we may lose information about how strongly the respondent *dislikes* certain features. That could slightly limit the model's ability to estimate utilities across the full range of preferences. And when ***k*** is close to ***n***, trimming becomes less impactful—there simply aren't many (or any) combinations without a selected item.

Image 14: Option 3A Coding of a 5 Choose 2 Task where Respondent Picked Item 2 and 4

Task	Alternative	Item #					Winner Winner	
		1	2	3	4	5		
1	1	0	1	0	1	0	1	← Winning Pair
1	2	1	1	0	0	0	0	
1	3	0	1	1	0	0	0	
1	4	0	1	0	0	1	0	
1	5	1	0	0	1	0	0	
1	6	0	0	1	1	0	0	
1	7	0	0	0	1	1	0	

Option 3B: Chunking by Chosen Features (With Added Tasks)

While Option 3A trims unchosen-only combinations, the remaining choice sets—especially when ***k*** is large—can still be quite large and complex. Option 3B builds on that idea but focuses on making the data more manageable for the model by breaking up the task into smaller, more focused pieces.

Instead of modeling one large choice set per task, Option 3B splits the data into separate "chunks," one for each chosen item. For example, if a respondent selects Features 2 and 4 out of 5, we create one choice set containing all combinations that include Feature 2, and a second set containing those that include Feature 4. Each chunk includes the original chosen combination plus other pairings of the selected feature with the non-chosen items.

In effect, this turns one 10-alternative task into two smaller, overlapping tasks. It's the same total number of alternatives overall—we're not adding or removing data—but we reorganize it so the model processes preferences in smaller blocks.

Image 15: Option 3B Coding of a 5 Choose 2 Task where Respondent Picked Item 2 and 4

Task	Alternative	Item #					Winner
		1	2	3	4	5	
2	1	0	1	0	1	0	1
2	2	1	1	0	0	0	0
2	3	0	1	1	0	0	0
2	4	0	1	0	0	1	0
3	1	0	1	0	1	0	1
3	5	1	0	0	1	0	0
3	6	0	0	1	1	0	0
3	7	0	0	0	1	1	0

All combinations with item 2 (the winner) (Alts 1-4 in task 1)

All combinations with item 4 (the winner) (Alts 1 and 5-7 in task 1)

Running New Models

After recoding our data to reflect each option, we re-estimated models and checked the holdout predictions. The table below summarizes the predictive accuracy (MAEs) for the baseline (Option 1) versus these new options. The green highlight is the best performer in the row and the red highlight is the worst performer in the row.

Image 16: In-Sample Mean Absolute Errors

	Dual Response Coding (Naïve None)	Option 1 Baseline	Option 2A Append Pairwise Comparisons	Option 2B Comparison-Only Tasks	Option 3A Trimming Unchosen Combinations	Option 3B Chunking by Chosen Features
Project A	Top Box	1.45	7.70	5.31	1.19	9.10
	70/30 Calibrated	3.16	10.89	5.61	3.11	14.27
	Top 2 Box	3.28	9.96	6.97	3.17	11.73
Project B	Top Box	4.46	11.00	8.42	4.14	12.34
	70/30 Calibrated	2.57	14.58	8.73	2.34	18.15
	Top 2 Box	1.80	7.60	4.98	1.62	8.73
Project C	Top Box	0.19	5.93	1.27	0.20	5.43
	70/30 Calibrated	0.42	9.53	1.33	0.56	10.28
	Top 2 Box	6.09	3.96	3.69	6.77	4.12

Image 17: Out-of-Sample Mean Absolute Errors

	Dual Response Coding (Naïve None)	Option 1 Baseline	Option 2A Append Pairwise Comparisons	Option 2B Comparison-Only Tasks	Option 3A Trimming Unchosen Combinations	Option 3B Chunking by Chosen Features
Project A	Top Box	3.33	7.66	5.32	3.27	9.03
	70/30 Calibrated	3.43	10.74	5.82	3.33	14.16
	Top 2 Box	5.80	9.81	7.61	5.76	11.57
Project B	Top Box	5.97	10.95	8.77	5.88	12.37
	70/30 Calibrated	4.32	14.28	8.85	4.00	17.92
	Top 2 Box	5.14	8.73	6.70	5.09	9.71
Project C	Top Box	1.62	5.49	2.06	1.47	5.34
	70/30 Calibrated	1.07	8.88	1.56	1.06	9.62
	Top 2 Box	6.77	4.48	3.90	7.66	4.70

The initial findings were intriguing. Option 3A (Trimming Unchosen Combinations) performed relatively well—its MAE was on par with or slightly better than the Option 1 baseline in some cases. Removing the "loser-only" alternatives didn't degrade performance and even seemed to help. This suggests that, at least in our data, little information was lost by cutting out combinations of unchosen features.

In contrast, Options 2A, 2B, and 3B underperformed. Their MAEs were significantly higher (worse) than the Option 1 baseline . This was true both in-sample and out-of-sample. In other words, simply adding a bunch of comparisons (2A) or splitting tasks (3B), without any other adjustments, actually hurt the model's ability to predict the holdouts. We also confirmed via log-likelihood analyses that these options were doing a poorer job at reproducing respondents' choices compared to the baseline.

At first glance, the results may seem counterintuitive. In Options 2A and 3B, we gave the model *more* data—so why did performance decline? And in 2B, we simplified the task structure—shouldn't reducing complexity help, or at least not hurt? Part of the answer might be in how hierarchical Bayes (HB) models interpret and weigh the data they're given.

Overfitting and Shrinkage

In an HB model, each respondent's part-worths are influenced by two forces: the individual's own data (likelihood) and the population distribution (prior). When a respondent has a lot more data points, the model trusts their data more and shrinks their utilities less toward the mean. In

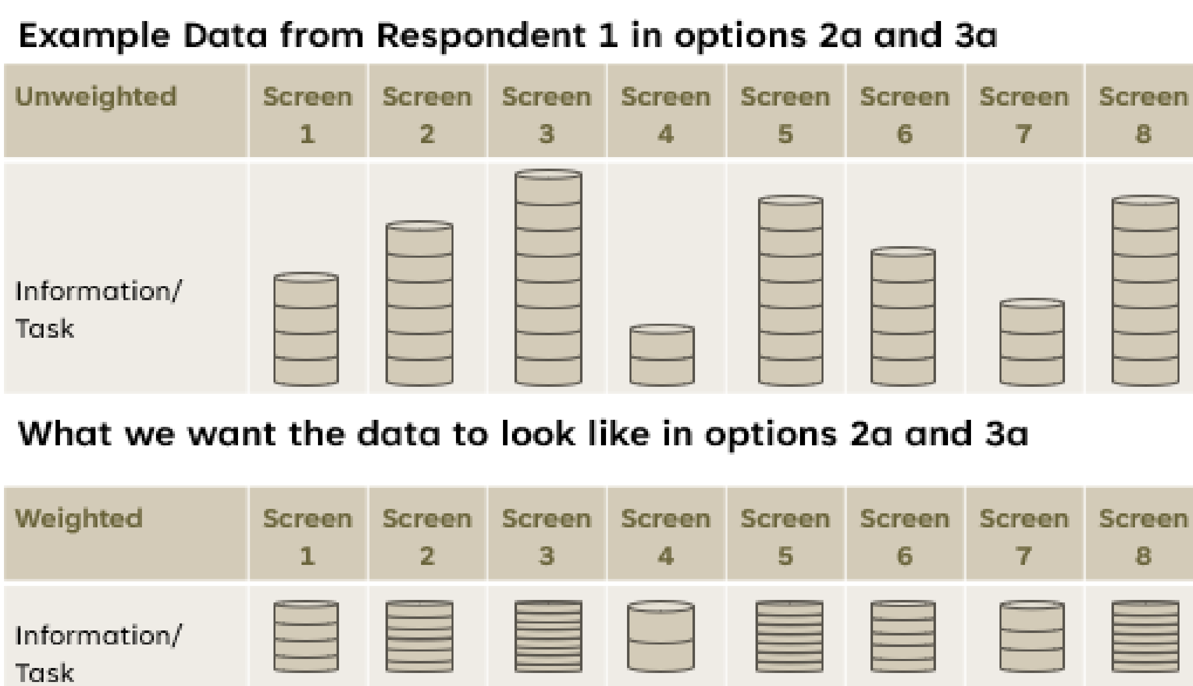
Options 2A, 2B, and 3B, each respondent's number of "observations" is effectively increased (because we added extra derived tasks for each of their original tasks). But those extra observations are not providing truly independent new information—they are derived from the same choice. The model doesn't know that, so it treats them like additional evidence and gives less weight to the prior. In short, we accidentally overfit each individual by feeding the model redundant data. The HB posterior for each person became more confident (less shrinkage), which can hurt out-of-sample prediction if those people had any noise in their choices (and everyone has some). This phenomenon can cause the overall fit (especially out-of-sample) to worsen, as we saw.

One challenge with Options 2A, 2B, and 3B is that they introduce imbalance in how much data each screen contributes, depending on the number of items selected (***k***). When ***k*** = *1*, only one comparison task is created. But when ***k*** = *6*, that same screen might produce six separate tasks—one for each selected item compared against the rest.

This means that screens with higher ***k*** generate significantly more data, even though it's not clear they provide proportionally more information. Choosing one item (***k*** = 1) reflects a clear top preference. But selecting six out of seven items (***k*** = 6) is closer to identifying the *least* important item—which may not be six times more informative. The result is an uneven data structure that can bias the model toward overfitting high-***k*** screens, while underweighting the sharper signals from low-***k*** ones.

Given these issues, we wondered whether applying a weighted likelihood could address both problems: the over-counting of data from added or chunked tasks, and the imbalance introduced by varying ***k***. By adjusting the weight of each task based on how it was generated, we aimed to restore balance and prevent any one screen from dominating the model.

Image 18: A Visual of the "Problem" of Imbalance with Low-k vs. High-k Screen



Weighted Analysis: Fixing the Balance

To address the imbalance, we modified our HB estimation to apply weighted likelihoods to the synthetic tasks. The goal was to ensure that each original screen contributes equally to the model, regardless of how many tasks were generated from it.

For example, if a screen produced ***k*** additional tasks (as in Option 2A), we assigned a weight of 1 / (***k***+1) to each of the ***k***+1 tasks—including the original. This kept the total contribution from that screen equal to 1, preserving parity across all screens.

Option 3A didn't introduce any new tasks; it only reduced the number of alternatives per task. As a result, no weighting adjustment was needed—its contribution remained balanced by design.

We re-estimated the models for Options 2A, 2B, and 3B using the weighted likelihood approach—and the results were striking. In-sample MAEs dropped significantly, in some cases even outperforming the Option 1 baseline. More importantly, out-of-sample MAEs improved across the board, matching or exceeding the baseline's performance.

Image 19: In-Sample Mean Absolute Errors with Weighted Likelihoods

	Dual Response Coding (Naïve None)	Option 1 Baseline	Option 3A Trimming Unchosen Combinations	Option 2A Append Pairwise Comparisons Weighted	Option 2B Comparison-Only Tasks Weighted	Option 3B Chunking by Chosen Features Weighted
Project A	Top Box	1.45	1.19	1.06	1.21	0.99
	70/30 Calibrated	3.16	3.11	2.96	3.10	3.16
	Top 2 Box	3.28	3.17	2.45	2.41	2.90
Project B	Top Box	4.46	4.14	1.80	1.61	3.47
	70/30 Calibrated	2.57	2.34	2.17	2.02	2.16
	Top 2 Box	1.80	1.62	0.71	0.62	1.12
Project C	Top Box	0.19	0.20	2.03	3.09	0.58
	70/30 Calibrated	0.42	0.56	0.25	0.26	0.45
	Top 2 Box	6.09	6.77	9.16	10.15	7.85

Image 20: Out-of-Sample Mean Absolute Errors with Weighted Likelihoods

	Dual Response Coding (Naïve None)	Option 1 Baseline	Option 3A Trimming Unchosen Combinations	Option 2A Append Pairwise Comparisons Weighted	Option 2B Comparison-Only Tasks Weighted	Option 3B Chunking by Chosen Features Weighted
Project A	Top Box	3.33	3.27	3.38	3.53	3.20
	70/30 Calibrated	3.43	3.33	3.75	4.02	3.52
	Top 2 Box	5.80	5.76	5.23	5.10	5.56
Project B	Top Box	5.97	5.88	5.40	5.32	5.64
	70/30 Calibrated	4.32	4.00	3.13	3.15	3.57
	Top 2 Box	5.14	5.09	4.58	4.39	4.89
Project C	Top Box	1.62	1.47	1.87	2.73	1.41
	70/30 Calibrated	1.07	1.06	1.18	1.02	1.13
	Top 2 Box	6.77	7.66	9.08	9.90	8.59

Notably, Option 2B—which had one of the worst unweighted prediction errors—became the best-performing model when weighted, delivering the lowest out-of-sample MAE among all conditions. Weighted versions of 2A and 3B also rebounded, performing on par with or slightly better than the baseline.

In short, weighting resolved the overfitting problem: it allowed the additional comparisons to enhance the model without overpowering it.

The likelihood-based validation supported these findings. Without weighting, models like Options 2B and 3B showed substantially lower log-likelihoods than the baseline, indicating they were less effective at ranking alternatives—especially in scenarios involving a None option. But once we applied weighted likelihoods, performance improved dramatically: the weighted versions of 2B and others matched or even exceeded the baseline, particularly in predicting purchase decisions.

Image 21: In-Sample Benefits Ranking Choice Sets: Sum(Log Likelihoods)/1000

		Naïve None					Diener None	
	DR Coding	**Opt 1**	**Opt 3A**	**Opt 2A weighted**	**Opt 2B weighted**	**Opt 3B weighted**	**Opt 1**	**Opt 1 Weighted**
Project A	T1B	-1.42	-1.41	-1.41	-1.41	-1.41	-1.42	-1.41
	Calibrated	-1.40	-1.39	-1.40	-1.41	-1.40	-1.40	-1.41
	T2B	-1.42	-1.41	-1.41	-1.41	-1.41	-1.42	-1.41
Project B	T1B	-1.38	-1.37	-1.37	-1.38	-1.37	-1.38	-1.37
	Calibrated	-1.36	-1.36	-1.38	-1.39	-1.36	-1.38	-1.38
	T2B	-1.38	-1.37	-1.37	-1.38	-1.37	-1.38	-1.37
Project C	T1B	-2.25	-2.25	-2.31	-2.34	-2.26	-2.25	-2.29
	Calibrated	-2.23	-2.25	-2.33	-2.36	-2.27	-2.31	-2.36
	T2B	-2.27	-2.26	-2.31	-2.34	-2.27	-2.27	-2.27

Image 22: In-Sample Naive None Choice Sets:
Sum(Log Likelihoods)/1000

		Naïve None					Diener None	
	DR Coding	Opt 1	Opt 3A	Opt 2A weighted	Opt 2B weighted	Opt 3B weighted	Opt 1	Opt 1 Weighted
Project A	T1B	-0.20	-0.19	-0.18	-0.18	-0.19	-0.34	-0.27
	Calibrated	-0.35	-0.34	-0.32	-0.32	-0.34	-0.70	-0.68
	T2B	-0.14	-0.14	-0.13	-0.13	-0.14	-0.26	-0.19
Project B	T1B	-0.21	-0.20	-0.18	-0.18	-0.20	-0.34	-0.27
	Calibrated	-0.37	-0.36	-0.34	-0.33	-0.35	-0.73	-0.72
	T2B	-0.13	-0.12	-0.12	-0.12	-0.12	-0.21	-0.16
Project C	T1B	-0.39	-0.37	-0.32	-0.31	-0.35	-0.62	-0.49
	Calibrated	-0.52	-0.50	-0.46	-0.45	-0.48	-1.01	-0.98
	T2B	-0.33	-0.31	-0.27	-0.27	-0.29	-0.52	-0.40

Conclusions and Key Learnings

Token-Based Conjoint (TBC) offers a practical and scalable solution to the long-standing challenge of evaluating large sets of product features. Across multiple studies, it proved effective at handling up to 40 attributes while maintaining both respondent engagement and analytical robustness.

Two innovations were important to TBC's success:

- The naïve none coding improved purchase prediction by consistently modeling the respondent's threshold across all tasks.
- Weighted likelihoods corrected for overrepresentation when tasks were added or chunked, preserving model balance and accuracy.

We also demonstrated that TBC can be extended beyond its original format. Option 3A (trimming unchosen-only alternatives) and Options 2B/3B with weighting (which split or simplify tasks) offer promising paths for scaling TBC to even more complex research designs.

When and Why to Use TBC

TBC is especially well-suited to studies with 20+ attributes, where traditional conjoint designs become too complex and MaxDiff lacks the additive structure required for bundle simulation. Compared to CBC, TBC focuses more on feature prioritization than direct trade-offs between complete product profiles. And unlike MaxDiff, it enables utility-based purchase simulations while still keeping the task lightweight for respondents.

From a respondent perspective, TBC was intuitive and engaging. Participants "built" their ideal product step-by-step and reported purchase likelihood, creating a more natural experience than completing back-to-back CBC and MaxDiff modules. From a stakeholder perspective, the outputs were straightforward and actionable: top feature rankings, uptake simulations, and clear diagnostics on how many features are needed to make a compelling offer.

However, TBC is not without limitations:

- Comparisons between bundle sizes (i.e., different values of k) are not directly observed. Any inferences across k rely on extrapolation via transitivity through the None response.
- Data is sparse for low-ranked features, so utility estimates for consistently unchosen items are less precise. Researchers may need to supplement TBC with targeted tasks or fixed evaluations if insights into bottom-tier features are critical.
- Interpretability can decline as modeling complexity increases. Techniques like synthetic task generation and weighting improve performance, but can make the model harder to explain to non-technical audiences.
- Performance beyond 40+ features remains untested. While early signs are promising, we have not yet fielded live studies with 60, 80, or more attributes.
- To date, no formal work has explored how to incorporate price sensitivity directly into token allocation or task design. So if price sensitivity and Willingness to Pay (WTP) analysis are key outputs for stakeholders, TBC is not the recommended approach. At best, one could consider including "$1 off your subscription" or "$5 off your subscription" as items within the TBC to get some signal on WTP.

Call to Action

TBC opens new possibilities for high-dimensional product research, but there is much to explore. As studies scale to 60+ features, adaptive designs—such as those inspired by Bandit MaxDiff—could be used to intelligently oversample features until their utility upper bounds are known.

Beyond scalability, future versions of TBC could benefit from modeling utility as a non-linear function of part-worths, rather than assuming simple additivity. This would allow for explicit diminishing returns as more items are added, better capturing how people experience bundles and helping prevent overstated uptake from purely additive models.

Last Words

In summary, TBC offers a powerful and respondent-friendly approach for understanding feature-level preferences in high-dimensional settings. By combining the strengths of MaxDiff and CBC with a simple yet informative token allocation mechanic and follow-up purchase intent, TBC enables deeper exploration without overwhelming respondents.

That said, TBC isn't universally the right tool. Its strengths shine when pricing isn't central—such as when prices are already fixed or the focus is on additive value. In contrast, if the aim is to include price as an attribute and optimize for revenue, we advise caution; methods like CBC or a MaxDiff + CBC hybrid may be more appropriate.

Ultimately, the choice of method should reflect what the client is trying to learn. Each approach comes with trade-offs. TBC has a valuable place in the toolkit—especially when tackling high-dimensional problems where traditional conjoint methods struggle. With thoughtful application, it can help teams prioritize features intelligently and design products that resonate.

Want to learn how to run a TBC from start to finish? Join our free TBC community within the Numerious Way at the link here: https://the-numerious-way.mn.co/share/tFITKjYYjpg0xh5B

Megan Peitz

Trevor Olsen

References

Brazell, J., C. Diener, E. Karniouchina, W. Moore, V. Séverin and P. Uldry 2006. The no-choice option and dual response choice design. *Marketing Letters,* 17:255–268.

Chrzan, K. and B. Orme. 2017. Becoming an expert in conjoint analysis. Sawtooth Software.

Diener, C., B. Orme and D. Yardley 2006. Dual response "none" approaches: theory and practice. In *Sawtooth Software Conference Proceedings,* pp. 157–168. Sequim, Wash.: Sawtooth Software.

Lattery, K. 2016. Dual response rating scale analysis. Paper presented at the Turbo Choice Modeling Event, Captiva Island, Fla.

COMMENT ON OLSEN AND PEITZ

DEAN TINDALL
SAWTOOTH SOFTWARE

Firstly, I want to commend the authors for demonstrating both the validity and practical value of this methodology. It's clear from the results that this approach offers real-world utility, particularly for clients looking to optimise their product tiering strategies—think "good-better-best" offerings. The rotation of features within each task is a clever design choice. It helps prevent respondents from simply picking the best value-for-money option every time (a common pitfall in fixed-price studies), and the inclusion of a None threshold provides a clear buy/no-buy decision at each level of feature inclusion.

BUILDING ON TURF ANALYSIS

I see this technique as a significant evolution from traditional TURF (Total Unduplicated Reach and Frequency) analysis. TURF is great for measuring the breadth of appeal—how many people like at least one of your offerings—but it often falls short on depth. In other words, just because a feature set reaches a lot of people doesn't mean it drives strong purchase intent. This new approach, by contrast, allows analysts to focus on combinations that not only reach but also actually convert customers. In my experience, many clients use TURF as a proxy for purchase intent, which can be misleading. This method, therefore, fills an important gap and could replace TURF in scenarios where depth of appeal is critical.

SUGGESTION: ZERO-FEATURE BASELINE

One simple but potentially powerful addition would be to include a task where the product has zero features—a true baseline. This would let us measure the inherent appeal of the product itself, without any added features. In the TURF chart shown, it looked like the first feature drove all the uplift, but that might be because the product already had some baseline appeal. Including this "no feature" scenario would help separate the value of the base product from the incremental value of added features. I'd encourage the authors to consider this for future iterations, as it could provide a more nuanced understanding of where the real value lies.

COMPARISONS WITH OTHER TECHNIQUES

This method sits in an interesting space alongside Q-Sort and MaxDiff. Like those techniques, it forces respondents to make tough trade-offs, but it does so in a way that's directly actionable for tiered product design. One area I'm curious about is the potential to integrate price as an additional parameter—something which will elevate this technique even further.

If the goal is to help clients design product tiers, it might make sense to tie price to the number of features (the "k" in each grouping), or to introduce price as a random element—perhaps even varying it between respondents or tasks. This could help capture the real-world tension between feature count and willingness to pay.

It would also be interesting to compare this technique head-to-head with a partial profile conjoint study, where the number of features per product is carefully controlled. This could help clarify where each method excels, and whether there are specific scenarios where one is clearly preferable.

DIMINISHING RETURNS

Finally, I'm not sure the current method fully accounts for diminishing returns as more features are added. In practice, we know that each additional feature tends to add less incremental value. It might be worth exploring whether the model could be adjusted to capture this effect—perhaps by explicitly modelling diminishing returns.

CONCLUSION

Overall, this is a thoughtful and practical methodology that builds on the strengths of existing approaches while addressing some of their key limitations. I'm excited to see how it evolves and would encourage further experimentation, especially around baseline measurement and price integration. I believe many clients would benefit from adopting this approach, particularly those looking to design product tiers which seek to optimise purchase intent within their target markets.

Dean Tindall

SYNTHETIC AI AVATARS IN MARKET RESEARCH—A GAME CHANGER OR A MERE GIMMICK?

SAURABH AGGARWAL
TARUN KHANNA
RASHMI SHARMA
KNOWLEDGE EXCEL

I. INTRODUCTION

Online surveys continue to drive market research, but as digital audiences grow more distracted and sceptical, response rates and data quality have suffered (Goodman, 2023). In an age of conversational interfaces like Siri, Alexa, and ChatGPT, the static, text-laden format of traditional questionnaires risks eroding respondent engagement. This widening engagement gap threatens both sample representativeness and the reliability of the insights on which brands depend.

Synthetic AI avatars, or "synthetic researchers" (Kuzmina et al., 2024), offer a potential remedy. By delivering survey questions through lifelike video agents that speak, gesture, and adapt to multiple languages, avatars may carry a promise to re-humanise the online interview while preserving perfect script standardisation.

This study contributes to the experiments that evaluate avatar-led surveys in an online environment. Moving beyond basic feasibility, we examine whether avatars genuinely elevate engagement, reduce behavioural sources of error (such as speeding and straight-lining), and maintain the analytical integrity of complex tasks like MaxDiff, all while remaining acceptable and appropriate from the respondent's point of view. In doing so, we aim to determine whether synthetic AI avatars represent a breakthrough for online market research or a passing novelty, and to outline practical guidelines for their responsible deployment at scale.

II. RESEARCH QUESTIONS AND APPROACH

This study asks not only whether synthetic AI avatars can boost engagement, but also which avatar design works best and why. Guided by the gaps noted in the introduction, we formulated the following research questions:

1. Can emerging technologies, such as AI avatars, influence respondent engagement, data quality, and overall survey experience across diverse use cases?
2. Is there a systematic preference for avatars of a particular gender, and does that preference influence reported engagement or overall survey experience differently for male and female respondents?
3. How do respondents generally perceive and accept the use of avatars in online surveys?

In addition to the questions above, the research aimed to integrate an AI avatar into a MaxDiff exercise and evaluate its impact.

Research Design

The study unfolded in three stages: Feasibility, Refinement, and Field Experiment, to move the avatars from concept to proof.

- Phase 1 confirmed technical viability (*Can we make this work?*).
- Phase 2 fine-tuned the prototype with expert feedback (*Can we make it better?*).
- Phase 3 deployed the optimised avatars with live respondents to quantify their impact (*Can they improve the survey experience in real-world conditions?*).

Phase 1: Feasibility Assessment

Commercial avatar platforms were stress-tested for voice naturalness, lip-sync accuracy, latency, and cross-device rendering. The outcome was a curated library of over 200 avatars across platforms, differing in age, gender, and ethnicity.

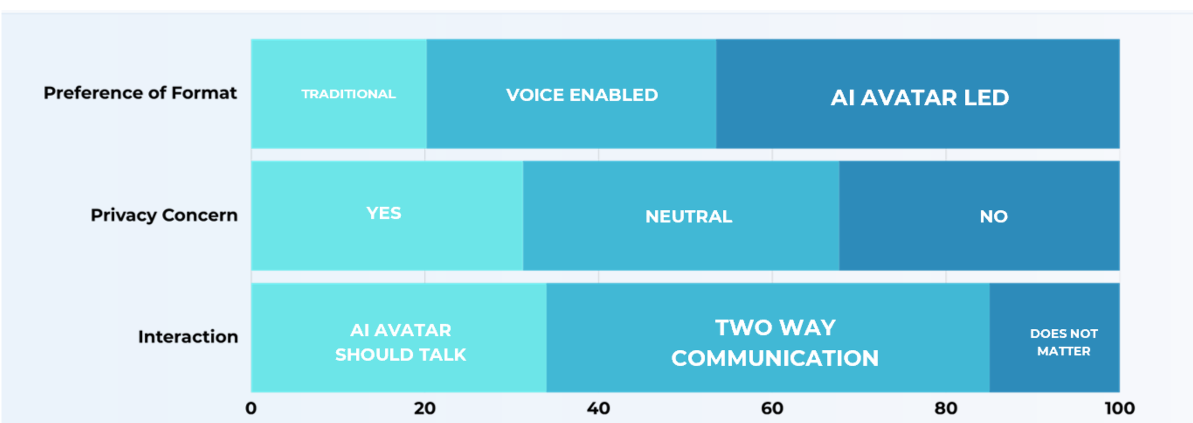
Phase 2: Prototype Validation with Research Professionals

In this phase of the research, the objective was to refine the avatar prototype before public rollout by obtaining rigorous, expert-level feedback on the concept.

A cohort of 187 market research professionals, recruited through LinkedIn posts, targeted e-mails, personal invitations, and follow-up one-to-one interviews, completed an eight-question survey narrated by a prototype avatar that replicated a typical respondent journey. Immediately afterward, they filled out a brief feedback form rating their experience of an AI Avatar led survey on the avatar's clarity, tone, pacing, and helpfulness versus intrusiveness, and provided open-ended suggestions for improvement.

This phase yielded three headline numbers: **69 %** said the avatar made the survey engaging, **58 %** felt comfortable using it, and **57 %** agreed it clarified the questions.

Figure 1: Research-Professional Feedback on Survey Format, Privacy, and Interaction Mode after the Phase 2 Avatar Walk-Through (N = 187)



When asked which survey format they would choose going forward, the plurality selected a fully AI-avatar-led survey, followed in size by a voice-only option and, last, a traditional text layout. Privacy worries were modest: the largest block sat in the neutral middle, with a smaller group expressing no concern and an even smaller group saying yes, they were worried. Finally, most respondents wanted richer interaction, about half preferred a two-way conversation with the avatar, roughly a third were happy if the avatar simply talked, and the rest felt the interaction mode doesn't matter (refer Figure 1).

Open-ended comments from the research professionals painted a balanced picture. On the plus side, many described the avatar as "natural and human-like," praising its ability to make a routine web survey feel "futuristic and interactive." Several noted that hearing the avatar walk through examples could give respondents a better grasp of complex tasks and saw clear potential for using avatars to deliver detailed instruction screens. At the same time, researchers highlighted a few friction points. First, they wanted explicit playback control, pause, replay, or skip, so fast readers could set their own pace. Second, they asked for the option to choose their own avatar, arguing that a single persona cannot suit every audience. Third, some experienced occasional *glitches or lag*, which disrupted the flow and broke immersion. Finally, many felt the narration speed was too slow compared with silent reading.

Reviewer comments became upgrades for the next phase. First, we embedded a play/pause/replay bar plus a "skip narration" link, giving fast readers control over pacing. To eliminate the buffering that some testers noticed, all video clips were shifted to server-side streaming. While a handful found the voice "robotic," acoustic re-balancing warmed the tone without altering its neutral persona. The strongest request, greater autonomy, led to two big changes: an avatar toggle that lets respondents switch the AI Avatar guide on or off at any moment, and an avatar-choice screen where they pick one of four personas. These gamified touches personalise the experience without imposing a single, prescriptive face. In short, every adjustment was designed around a simple principle: meaningful human connection comes from giving respondents more choice and control, not from pushing harder.

Phase 3: Field Experiment ("Reality Check")

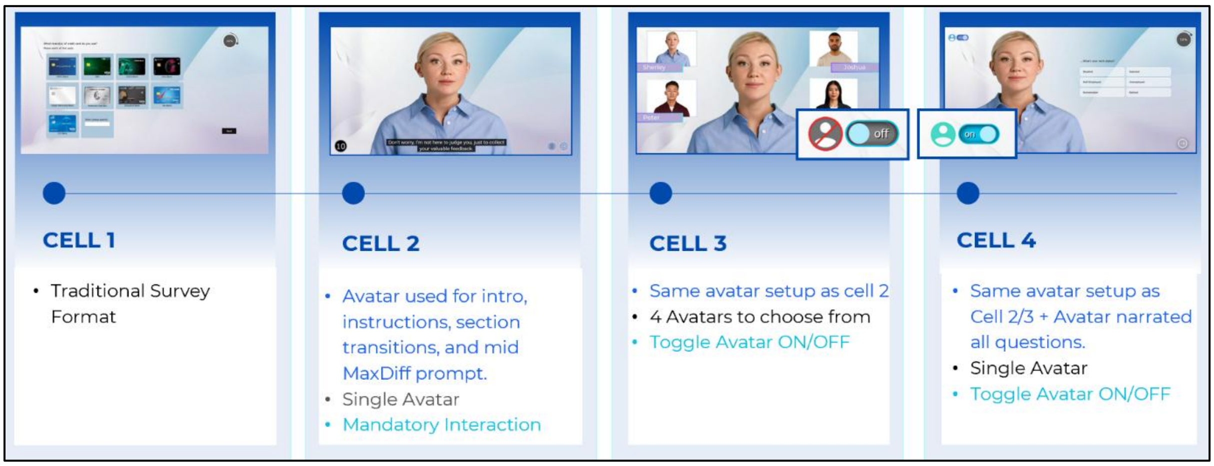
In this phase, we tested the AI-avatar solution under live, online conditions, asking: Does an avatar-led interface outperform a traditional text survey, and, if so, at what level of exposure and respondent control?

Sample

- N = 1668 adults, 18–55 years, recruited from panels, in India's metro cities
- Category focus: Consumer decision-making on credit-card features
- Devices: Mobile and desktop
- Fieldwork was conducted in March–April 2025

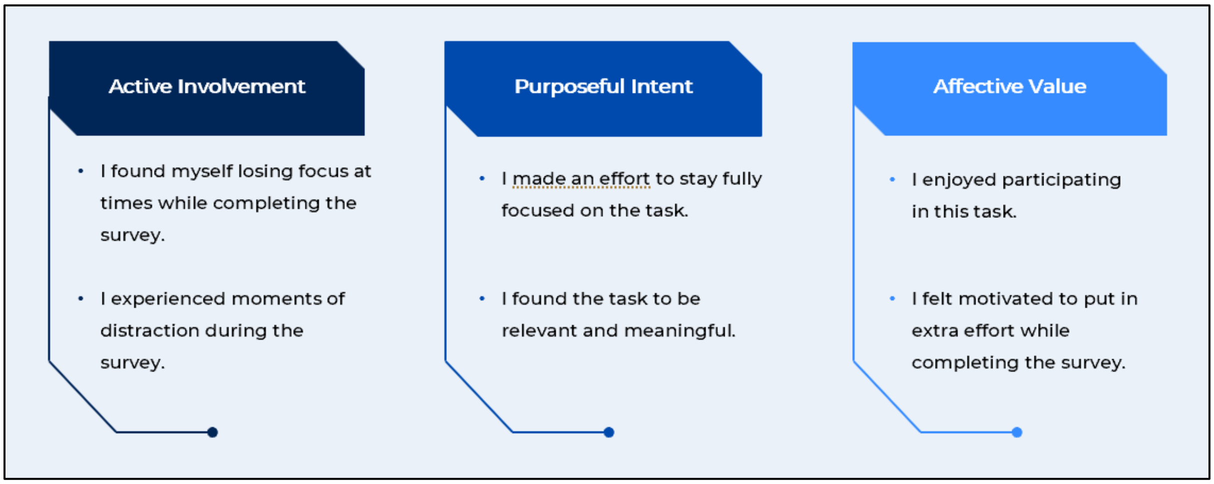
The sample was evenly split to take the same questionnaire presented in four distinct formats, each adding a new layer of avatar involvement. Cell 1 (N=468) formed the control group (baseline): a purely text-based questionnaire with no avatar involvement. The remaining three cells formed the test group. Cell 2 (N=400) introduced a female AI avatar that appeared at key junctures: welcome introduction, each section break, the MaxDiff briefing, and a mid-MaxDiff exercise progress nudge, and on submission page, while all clips carried the standard play, pause, and replay bar so respondents could control pacing. Cell 3 (N=400) layered personal choice and greater autonomy onto that same structure: participants first selected one of four avatars (two male, two female) and could toggle narration on or off at any moment, in addition to using the playback bar. Building on Cells 2 and 3, Cell 4 (N=400) provided continuous exposure, with the avatar present on every page, narrating each question by default. The toggle on/off and playback controls remained available, so respondents still dictated the presence of the avatar on the screen. Together, these four conditions let us test how varying the dose of avatar presence and the degree of user control influences engagement, data quality, and overall survey experience (refer Figure 2).

Figure 2: Phase 3 Experimental Cell Configurations



After completing the survey, all participants provided brief feedback on their experience. The control group answered a single overall-experience question, while the test group responded to that same item plus six additional questions about their interaction with the AI Avatar. The first feedback question captured the overall survey experience of the respondent with six items adapted from the Hannum and Simons Engagement Questionnaire (Hannum and Simons, 2020), each rated on a 5-point scale (1 = Strongly Disagree, 5 = Strongly Agree). Following the categorization outlined by Kuzmina et al. (2024), engagement attributes were grouped under three dimensions - Active Involvement, Purposeful Intent, and Affective Value (refer Figure 3).

Figure 3: Adapted Hannum and Simmons Engagement Questionnaire; Six Statements Mapped to Three Themes—Active Involvement, Purposeful Intent, and Affective Value



The next six feedback questions centred on the AI-avatar experience (only answered by the test group). First, respondents rated three statements, whether the avatar made the survey more engaging, increased their comfort, and helped them understand the questions, on the same 5-point *Strongly Disagree → Strongly Agree* scale used in Phase 2, enabling a direct comparison with the earlier researcher feedback.

Second, to test for a possible uncanny-valley reaction, the uneasy feeling people report when a digital face looks almost, but not quite, human, we asked respondents to rate the avatar on six bipolar adjective pairs: Machine-like vs. Human-like, Monotonous vs. Engaging, Friendly vs. Unfriendly, Confusing vs. Clear, Complicated vs. Simple, and Impersonal vs. Relatable. Large skews toward the negative poles would signal that the avatar had slipped into the uncanny valley; balanced or positive ratings would suggest an acceptably human-machine blend.

The next three questions were to capture whether respondents would be willing to take part in future surveys that use an AI avatar, the perceived appropriateness of the AI Avatar presence in the survey, and privacy concerns to see whether the presence of an avatar heightened data-security anxieties. Finally, an open-ended prompt invited respondents to offer suggestions or comments on how the avatar experience could be made more engaging or helpful, giving qualitative depth to the quantitative findings.

We thank InnovateMR and Takumi International for their support in fielding and data collection.

III. Results

The data were analysed across five primary domains to evaluate the impact of the AI Avatar on respondent experience and data outcomes: engagement levels, task performance within the MaxDiff exercise, overall data quality, presence of social preference bias, and respondent perceptions of the avatar. The findings for each domain are presented in the following sections.

A. Engagement Check: Did the AI Avatar improve respondent experience beyond traditional formats?

Length of Interview

The average survey duration increased with the level of AI Avatar presence across the four experimental cells (refer Table 1). Cell 1, which included no avatar interaction, recorded the shortest interview time (9.95 minutes average). In contrast, Cell 4, where the avatar was present throughout the survey, recorded the longest average duration (18.75 minutes average).

This pattern was consistent across device types. On mobile devices, interview lengths ranged from 8.70 minutes (Cell 1) to 17.00 minutes (Cell 4), while on desktop devices, durations ranged from 11.20 minutes (Cell 1) to 20.10 minutes (Cell 4). The extended duration in AI Avatar cells reflects additional interaction time associated with avatar-led delivery of instructions and questions.

Table 1: Average Interview Length (Minutes) Across Experimental Cells, Overall and by Device Type

Format	Cell 1	Cell 2	Cell 3	Cell 4
Overall	9.95	13.75	12.00	**18.75**
Mobile	8.70	13.20	12.10	**17.00**
Desktop	11.20	14.30	13.90	**20.10**

Completion Rate

Completion rates varied across the experimental cells, showing a decline as avatar presence increased (refer Table 2). Cell 1, which had no avatar, recorded the highest completion rate (91%). Cells with partial avatar involvement, Cell 2 (78%) and Cell 3 (82%), showed moderate declines. The lowest completion rate was observed in Cell 4 (68%), where the avatar was present throughout the survey. The reduction in completion rates with increasing avatar exposure may indicate added time or effort required when interacting with the avatar interface.

Table 2: Completion Rates Across Experimental Cells

Metric	Cell 1	Cell 2	Cell 3	Cell 4
Completion Rate	91%	78%	82%	**68%**

Engagement Scale

The engagement scale results indicate that moderate levels of avatar presence (Cells 2 and 3) were associated with stronger engagement across both cognitive and affective dimensions. Cell 2 appeared to offer the most emotionally fulfilling experience, while Cell 3 provided the strongest cognitive engagement. In contrast, Cell 1, with no avatar, demonstrated decent focus and a generally positive emotional experience, though at lower levels than Cells 2 and 3. Full avatar exposure in Cell 4, while maintaining relatively strong emotional engagement, showed some decline in cognitive involvement, suggesting potential overstimulation or cognitive fatigue. These patterns highlight the importance of calibrating avatar exposure to balance emotional satisfaction with sustained cognitive engagement. All engagement scores were calculated using Top 2 Box percentages, with reverse coding applied where necessary.

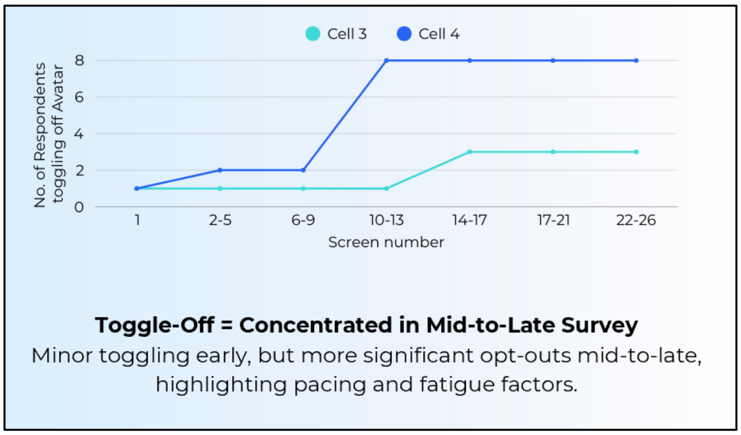
Table 3: Engagement Scale Results (Top 2 Box %) Across Experimental Cells (Adapted from Hannum and Simmons Engagement Questionnaire)

Theme	Cell 1	Cell 2	Cell 3	Cell 4
Active Involvement	64%	73%	**75%**	62%
Purposeful Intent	82%	85%	**88%**	87%
Affective Value	79%	**95%**	91%	83%

Avatar Toggle Behaviour

Engagement was further evaluated through respondents' voluntary use of the avatar toggle feature, available in Cells 3 and 4.

Figure 4: When did they toggle off?



Most respondents (93.5%) chose to keep the avatar activated throughout the survey. Among those who opted to disable the avatar, most did so during the mid-to-late stages of the survey, particularly after screen 10 (refer Figure 4). Early opt-outs were minimal. This pattern suggests that while initial engagement with the avatar was high, prolonged exposure may contribute to pacing or fatigue concerns for some respondents. These findings point to the potential benefit of introducing adaptive or flexible avatar configurations in longer surveys to sustain engagement while minimizing respondent burden.

B. Maxdiff Performance: Did responses become more thoughtful, consistent, and reliable with Avatar assistance?

MaxDiff was deliberately selected for this experiment due to its straightforward yet cognitively engaging design that offered a sensitive testbed for evaluating subtle effects of AI Avatar integration on response consistency and model fit. Any beneficial or disruptive impacts identified in this context could indicate broader implications for avatar use in more cognitively demanding methodologies, such as conjoint analysis.

The AI Avatar was introduced at two targeted points during the MaxDiff task: (a) during an information acceleration screen to explain task instructions, and (b) through a mid-exercise progress prompt to sustain engagement.

Two complementary indicators were used to evaluate data quality: Root Likelihood (RLH) and Mean Absolute Error (MAE). While higher RLH values indicate stronger internal consistency, lower MAE values indicate better model fit.

Table 4: Impact of Avatar Usage on MaxDiff Consistency (RLH) and Accuracy (MAE)

Cell	Cell 1	Cell 2	Cell 3	Cell 4
RLH[1]	0.378	0.390	0.384	0.366
Interpretation (RLH)	Baseline performance	Highest consistency: minimal, well-placed guidance	Nearly matched Cell 2: control improved clarity	Lowest consistency: cognitive overload risk
MAE[2]	0.013	0.018	0.016	0.015
Interpretation (MAE)	Baseline: excellent alignment between fixed & random tasks	Slightly higher, but still excellent: instruction-level avatar use didn't disrupt consistency	Personalization & control maintained strong utility alignment	Full avatar presence showed no meaningful distortion

Response consistency, measured by Root Likelihood (RLH), improved in conditions featuring either minimal or respondent-controlled avatar interactions relative to the baseline without an avatar. In contrast, consistency declined under continuous avatar narration, suggesting a potential increase in cognitive load when avatars were persistently present in the survey. However, predictive accuracy, assessed by Mean Absolute Error (MAE), remained uniformly strong across all experimental cells, clearly indicating that even extensive avatar usage did not disrupt overall model fit (refer Table 4).

These findings suggest avatars, when used strategically, can enhance respondent clarity and consistency without negatively impacting predictive accuracy. Conversely, continuous avatar use may increase cognitive demands without yielding additional benefits in prediction quality.

C. Data Quality: Did responses become more thoughtful, consistent, and reliable with Avatar assistance?

Open-End Responses

To extend the evaluation of data quality beyond structured choice tasks, open-ended responses were analysed for response richness and cognitive effort.

[1] Higher RLH value implies more consistency.
[2] Lower value of MAE implies better model predictiveness

Table 5: Open-Ended Response Category Distribution by Experimental Cell

Cell	Cell 1	Cell 2	Cell 3	Cell 4
Noise (1–2 words)	13%	6%	9%	5%
Short (3–4 words)	16%	9%	7%	11%
Acceptable (5–8 words)	22%	25%	27%	20%
Thoughtful (8+ words)	49%	60%	57%	64%

Responses were categorized into four tiers based on word count and interpretability: Noise (1–2 words, e.g., “good,” “fine”), Short Response (3–4 words; borderline, requiring contextual review), Acceptable (5–8 words; generally sufficient for interpretation), and Thoughtful (8+ words; detailed enough for qualitative analysis).

While differences across experimental cells were not large, a directional pattern was observed (refer Table 5). Cells incorporating AI Avatar assistance (Cells 2, 3, and 4) yielded a lower proportion of noise-level responses and a higher share of thoughtful, detailed answers compared to the baseline text-only condition (Cell 1). Notably, Cell 4, where the avatar remained present throughout, recorded the highest proportion of thoughtful responses. This suggests that avatar-led interaction may foster greater cognitive engagement, encouraging respondents to provide more elaborated feedback. This underscores AI Avatar’s utility in generating richer qualitative datasets, thereby enabling more robust thematic exploration and strengthening the validity of subsequent interpretive analyses.

Respondent Behaviour

Respondent behaviour analyses revealed a clear, inverse relationship between AI Avatar involvement and the prevalence of straightlining[3] and speeding[4]. In the absence of an avatar, both behaviours were most pronounced, indicating a higher likelihood of perfunctory responding (refer Table 6). Introducing even minimal avatar prompts led to a substantial decrease in these undesirable patterns. Granting respondents control over avatar exposure preserved this improvement, while continuous avatar presence achieved the greatest reduction in straightlining and speeding. These findings underscore the capacity of AI Avatars to mitigate low-effort response shortcuts, thereby bolstering the integrity and reliability of survey data.

Table 6: Impact of AI Avatar Presence on Straightlining and Speeding Rates

Metric	Cell 1	Cell 2	Cell 3	Cell 4
Straightliners	8%	2%	1.9%	1.1%
Speeders	12%	1.5%	2%	0.8%

[3] Anyone who gave the same response on ≥ 80 % of items within all the grid questions across the survey was considered as a speeder for this research.

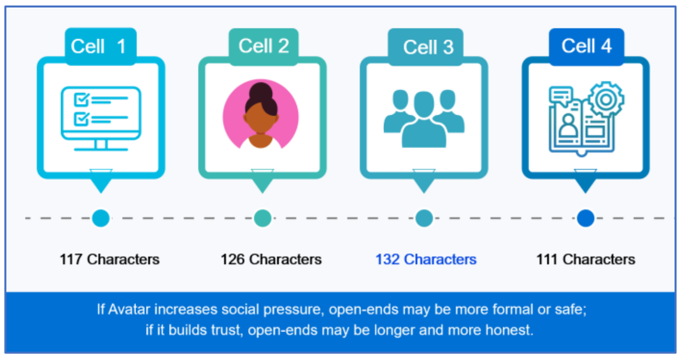
[4] Anyone whose total survey completion time is < ⅓ of the survey’s median time, was considered as a straightliner for this research.

D. Social Preference Bias: Were male or female avatars perceived differently by respondents?

To explore whether respondents exhibited a subconscious preference for male or female avatars, and how that choice affected their comfort and expressiveness, we analysed four key indicators: open-ended response depth, avatar selection rates, engagement experience, and overall survey experience.

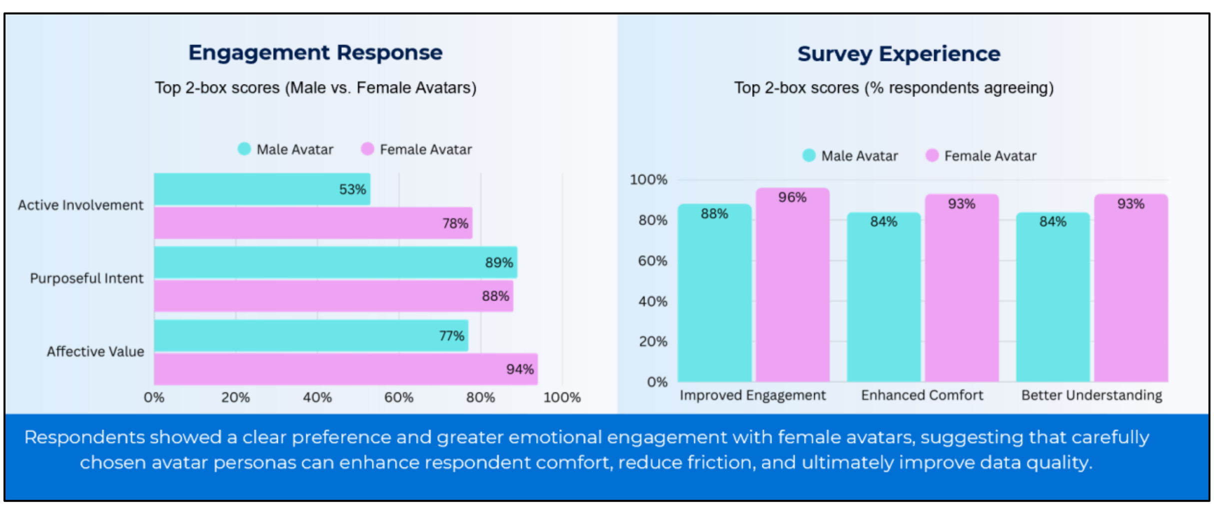
First, overall open-ended response length demonstrated that interaction autonomy drove expressiveness more than gender. The toggle-control condition (Cell 3) produced the longest verbatim feedback, whereas continuous avatar presence (Cell 4) saw a decline in response depth (refer Figure 5).

Figure 5: Average Open-End Verbatim Count across the Four Cells



When respondents were offered a choice of avatar gender in Cell 3, 81% selected the female persona. This strong preference likely reflects perceived approachability and lower social scrutiny associated with a female avatar, which may encourage more honest and uninhibited feedback in sensitive contexts.

Figure 6: Top 2-Box Engagement and Survey Experience Scores by Avatar Gender, for Cell 3



Further analysis revealed that female avatars not only were chosen more often but also elicited higher engagement ratings (refer Figure 3 for the six statements mapped to three themes—Active Involvement, Purposeful Intent, and Affective Value, used to calculate engagement responses metric) and clearer survey experiences than male avatars (refer Figure 6). Respondents reported feeling more actively engaged, comfortable, and confident in their understanding of questions when interacting with a female persona.

These findings indicate that while providing interaction control is critical for maximizing open-ended engagement, avatar gender can further shape respondent experience. Deploying a female avatar may enhance emotional engagement and, consequently, the richness and reliability of qualitative and quantitative data in survey research.

E. Avatar Perception: How did respondents feel about interacting with the Avatar?

Respondents' attitudes toward the AI Avatar were assessed across six dimensions: engagement, perceived appropriateness, privacy concern, intent for future participation, uncanny valley effect, and open-ended feedback.

Table 7: Comparative Engagement Metrics Across Phase II Pilot and Phase III Avatar Conditions

	PHASE II	PHASE III		
Top 2 Box Agreement		**Cell 2**	**Cell 3**	**Cell 4**
Improved Engagement	69%	89%	99%	92%
Enhanced Comfort	58%	88%	96%	90%
Better Understanding	57%	91%	96%	92%
Engagement Score	**61.3%**	**91.7%**	**97%**	**92%**

Engagement rose markedly from the researcher pilot to the respondent sample, with all avatar conditions exceeding 90% on the Top 2 Box metric (refer Table 7). The toggle-control condition delivered the highest engagement, outperforming both instructional-only and continuous-presence setups.

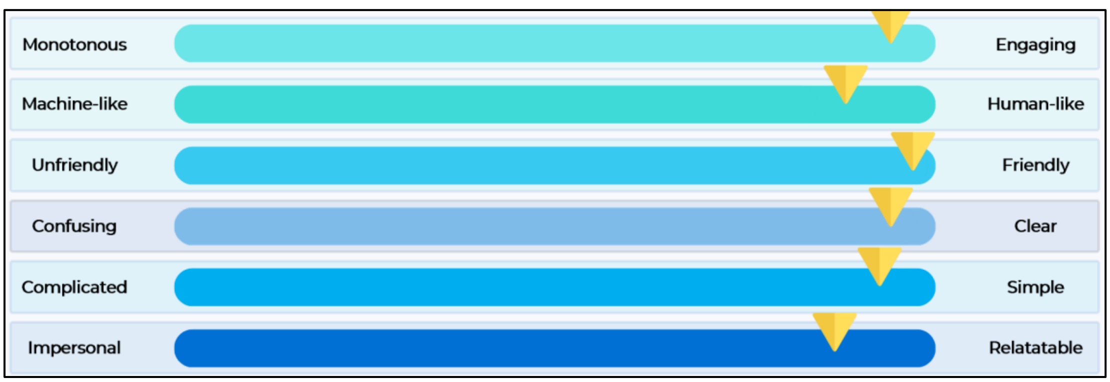
Respondents showed near-universal willingness to join future avatar-led surveys, with only a slight drop under continuous exposure. Perceptions towards appropriateness of the use of the Avatar in the survey remained high, though "perfect implementation" endorsements reduced slightly as avatar use intensified. Privacy concerns rose modestly with greater avatar presence but were still outweighed by neutral or no-concern responses (refer Table 8).

Table 8: Respondent Future Participation, Appropriateness Ratings, and Privacy Concerns by Avatar Exposure Condition

	Cell 2	Cell 3	Cell 4
Willingness to Participate in AI Avatar surveys in future (5 point rating scale question)			
Top 2 Boxes Agreement	98.2%	97.1%	93%
Perceived Appropriateness of AI Avatar Use in the Survey (single select question)			
Overused	9%	9%	11%
Underused	4%	6%	9%
Perfectly Implemented	89%	85%	80%
Privacy Concern (single select question)			
Yes	26%	30%	37%
Neutral	32%	21%	16%
No	42%	49%	44%

To ensure that our AI Avatar enhanced rather than undermined the respondent experience, we evaluated whether its design risked falling into the "uncanny valley," a phenomenon where near-human likeness can evoke discomfort.

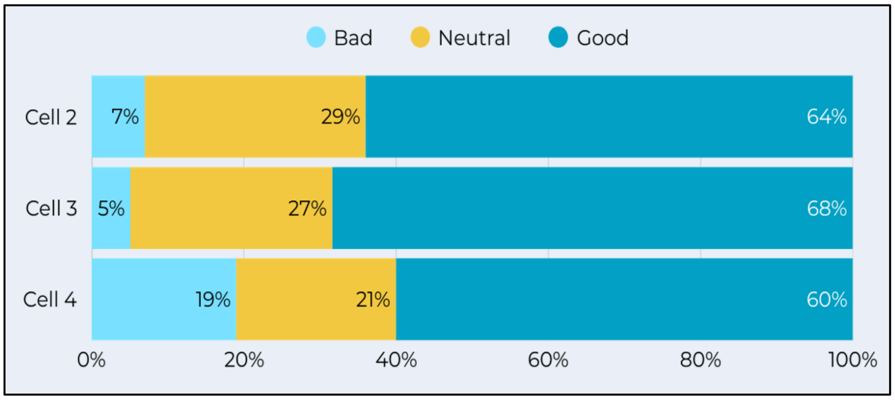
Figure 7: Mean Scores of Participant Perceptions of Avatar Human-Likeness and Approachability



All six semantic-differential dimensions skewed strongly toward the positive poles (refer Figure 7), 94% respondents saw the avatar as engaging (not monotonous), 86% of the respondents saw the avatar as human-like (not machine-like), 97% of the respondents saw the avatar as friendly (not unfriendly), 92% of the respondents saw the avatar as clear (not confusing), 89% of the respondents saw the avatar as simple (not complicated), and 82% of the respondents saw the avatar as relatable (not impersonal). This overwhelmingly positive profile indicates that the avatar did not trigger any uncanny-valley discomfort and was perceived as natural and approachable throughout the survey.

Finally, beyond numeric ratings, respondents were invited to share their thoughts on the avatar experience in an open-ended question. Open-ended feedback sentiment remained overwhelmingly positive (refer Figure 8), around 60%, under both instructional-only (Cell 2) and toggle-control conditions (Cell 3), with neutral ("indifferent") responses decreasing when avatars were continuously present (Cell 4). Conversely, negative feedback rose sharply, tripling to 19%, in the continuous-presence condition, indicating respondent fatigue from continuous avatar interaction. These patterns underscore that while AI Avatars generally enhance qualitative engagement, sustained exposure without user control may provoke fatigue and annoyance, highlighting the need for calibrated, respondent-driven avatar deployments.

Figure 8: Open-End Response Sentiments towards AI Avatar



IV. Practical Guidelines for AI Avatar Deployment

Drawing on our empirical research and casual empiricism, we present a cohesive framework with five guiding principles, each elaborated through targeted practices and applications.

A. Transparency

Inform respondents at the outset that an AI Avatar will guide the survey. Explain its purpose and reaffirm data privacy and security measures to establish trust and set clear expectations.

B. Design and Localization

Visual and Vocal Design: Use a neutral yet friendly avatar with subtle facial cues and a warm, steady voice. Language should be inclusive and jargon-free.

Cultural Adaptation: Localize gestures, idioms, and persona traits to resonate with respondents' cultural contexts, avoiding generic "stock" characters.

C. Strategic Deployment and Autonomy

Targeted Presence: Integrate avatars at cognitively demanding moments, survey introductions, complex exercise explanations (e.g., MaxDiff, conjoint), and mid-survey prompts, instead of continuous narration.

User Control: Provide on/off toggles and optional persona selection to empower respondents. This autonomy maintains engagement and reduces fatigue.

Error and Progress Prompts: Leverage avatars to nudge for missing responses and to celebrate milestones or track incentives in lengthy or multi-stage surveys, giving respondents a constructive company though the survey.

Debrief and Feedback: One may conclude with a brief avatar-led thank-you and request final comments to foster goodwill and collect qualitative insights.

D. Technical Performance

Adaptive Streaming: For implementation of AI Avatar videos in surveys, stream video segments online to prevent lag across bandwidths.

Playback Controls: On platforms that block autoplay (iOS/macOS), include visible play buttons or use single-page scripting to trigger avatar segments.

Offline Modes: Since streaming would not work in an offline environment, preload all avatar assets locally to ensure smooth playback.

Cross-Platform Testing: Validate performance across devices, browsers, and connection speeds to anticipate and resolve rendering issues before field deployment.

E. Suitability and Applications

Cautionary Contexts: Avoid extensive avatar use in repetitive choice tasks (all MaxDiff/ Conjoint tasks) and culturally nuanced dialogues requiring human intuition.

By following these principles, researchers can implement AI Avatars that enhance clarity, boost engagement, and protect data integrity while respecting respondent autonomy and comfort.

V. Future Vision: Intelligent AI Agents

The developments reported in this study represent only the first phase of AI Avatar integration. Current engineering efforts are focused on evolving these avatars into fully intelligent agents powered by large language models (LLMs). Such agents will transcend scripted interactions by engaging respondents in spontaneous, two-way dialogues, enabling dynamic probing, contextual clarifications, and on-demand guidance during survey administration.

Key capabilities envisioned for these intelligent agents include:

- **Natural Language Understanding and Generation:** Agents will interpret open-ended replies, ask follow-up questions in real time, and adapt phrasing based on respondents' linguistic styles and comprehension levels.
- **Capture Qualitative, Populate Quantitative:** A hands-free, two-way interactive survey experience where the AI Avatar would ask questions verbally, and respondents' answers would be captured verbatim using speech-to-text technology. These responses would then be integrated into a quantitative survey, auto filling the necessary fields, thus bridging the gap between qualitative insights and quantitative data collection.
- **Knowledge Bot Integration:** The avatar will serve as the front end for domain-specific knowledge bots, allowing respondents to pose contextual questions and receive immediate, informed responses without leaving the survey interface.

The transition to intelligent AI agents promises to redefine survey methodologies by blending conversational richness with rigorous data capture. Future research should evaluate the reliability, validity, and ethical implications of such systems, including the impact on respondent experience, respondent behaviour and data quality. By integrating LLM-driven agents, researchers will be able to explore new avenues for immersive, adaptive, and highly personalized survey experiences for data collection.

VI. Conclusion

This study illustrates that integrating AI Avatars into surveys can meaningfully enrich respondent engagement and enhance data quality. Thoughtfully timed avatar interventions facilitate deeper cognitive and emotional involvement, reduce low-effort response behaviours, and improve the consistency and accuracy of choice tasks. Qualitative feedback further confirms that avatars clarify instructions and cultivate a more personable survey environment without triggering discomfort or bias.

These benefits, however, come with trade-offs. Avatar use tends to lengthen interviews, may require higher incentives to sustain motivation, and can modestly depress completion rates. The key insight is that richer engagement often justifies these costs, particularly when survey objectives prioritize depth of insight and data integrity over the speed of data collection alone.

Ultimately, researchers must balance these considerations considering their specific goals. By leveraging the Practical Handbook's guidelines (elaborated in section IV of this research paper) and anticipating next-generation, LLM-powered agents, practitioners can tailor avatar deployment to achieve the optimal blend of engagement, efficiency, and data quality.

Saurabh Aggarwal

Tarun Khanna

Rashmi Sharma

References

DaveAI (2024, August 19). *Businesses adopting avatar-based surveys saw a 30% increase in survey engagement*. LinkedIn. https://www.linkedin.com/pulse/businesses-adopting-avatar-based-surveys-saw-30-increase-survey-ohyof/

Goodman, J. (2023). *Understanding the Decline in Survey Participation: Challenges and Solutions*. Survey Research Journal, 45(2), 115–130

Hannum, M. E., & Simons, C. T. (2020). Development of the Engagement Questionnaire (EQ): A tool to measure panelist engagement during sensory and consumer evaluations. *Food Quality and Preference, 81*, 103840. https://doi.org/10.1016/j.foodqual.2019.103840

Kuzmina, A., Sismey, M., Imandar, M., Do, B., & Delahunty, C. (2024). *Can synthetic AI avatars help increase survey engagement and deliver better data quality in Asia?* ESOMAR. https://ana.esomar.org/documents/can-synthetic-ai-avatars-help-increase-survey-engagement-and-deliver-better-data-quality-in-asia

Miao, F., Kozlenkova, I. V., Wang, H., Xie, T., & Palmatier, R. W. (2021). An emerging theory of avatar marketing. *Journal of Marketing*, *85*(4), 64–86. https://doi.org/10.1177/0022242921996646

TRANSFORMING OPEN-ENDED SURVEY RESPONSE ANALYSIS WITH AI TRANSFORMER MODELS

JACOB NELSON
THE HARRIS POLL

INTRODUCTION

Open-ended questions are a staple in survey research for good reason. They give respondents the freedom to answer in their own words, often revealing richer, more nuanced insights than closed-ended questions can capture. Open-ends allow researchers to hear the language people actually use, detect emotional undertones, uncover unmet needs, and even indicate survey fraud. They're especially useful when exploring new categories, testing early concepts, or trying to understand why respondents feel the way they do.

But for all their benefits, open-ended questions come with a cost. Once the data is collected, someone has to make sense of the raw text, and that process is almost always slow, messy, and manual. Traditional coding requires teams of analysts to read, tag, and interpret every response. It's time-consuming, expensive, and often inconsistent. As a result, many researchers default to simpler yes/no or scale-based questions. Not necessarily because they yield better insights, but because they're easier to analyze at scale.

As in many domains, both within market research and beyond, *AI models are changing the game*. This paper explores two distinct but complementary tools for AI-assisted coding: one that uses an embeddings model to group responses based on semantic similarity, and another that applies large language models (LLMs) to reason about the structure and taxonomy of responses directly. Both rely on the same underlying transformer model architecture, but they apply it in different ways. We show how each method can help researchers scale up their open-end analysis without sacrificing depth or quality, and make it more feasible to include more open-ended questions in market research.

RELATED WORK AND BACKGROUND

Historically, researchers have relied on manual coding or simple statistical methods to analyze open-ended survey responses. Manual coding, while often effective at capturing nuance, is time-consuming, labor-intensive, and prone to inconsistencies across coders or studies. As a result, open-ended data is often underutilized in large-scale research, despite its potential for capturing rich, emergent insights.

To ease this burden, a number of computational approaches have been developed over the past two decades. One of the earliest was TF-IDF (Term Frequency–Inverse Document Frequency), which scores the importance of words based on their frequency within and across documents. TF-IDF can be useful for basic keyword extraction, but lacks the ability to account for context or semantics—a critical limitation when analyzing short, variable-length responses like those found in most surveys. Still, techniques based on TF-IDF and related clustering algorithms continue to offer value in some settings. For instance, Adrogue (2023) demonstrated how such methods can be applied to quantify open-ended speech in specific research contexts.

Later methods, like Latent Dirichlet Allocation (LDA) (Blei et al., 2003), aimed to uncover latent themes by modeling patterns of word co-occurrence. LDA has seen broad use in academic and marketing contexts but is often difficult to tune and interpret when applied to short, variable-length responses like those found in surveys. Because these models rely heavily on word frequencies, they also struggle to group responses that express similar ideas using different language.

More recently, transformer-based models have made it possible to interpret text with much greater nuance and contextual awareness. These models underpin tools like embeddings and LLMs, which are now widely used in modern NLP tasks, including classification, clustering, and summarization. This paper builds on those developments to examine how transformer-based approaches can improve the accuracy, scalability, and usability of open-end analysis in a market research setting.

THE TRANSFORMER ARCHITECTURE

The two AI methods explored in this paper—embeddings and large language models—are both built on the same underlying machine learning foundation: the transformer architecture. First introduced by Vaswani et al. (2017), transformers have become the standard approach in NLP due to their ability to interpret text with context, nuance, and long-range dependencies.

At the heart of the transformer is its self-attention mechanism, which allows the model to consider how each word in a sentence relates to every other word. This enables the model to infer meaning not just from isolated terms, but from how they function together within broader linguistic context. In contrast, older techniques like TF-IDF or bag-of-words models treat words independently, often missing the intent behind how something is said.

For example, consider the phrase: *"I don't need all that other jazz."* A frequency-based model might interpret "jazz" literally, perhaps inferring a reference to music. A transformer, by contrast, can infer from context that "jazz" here means "unnecessary extras," capturing the respondent's intent rather than just the surface language.

This ability to parse meaning and detect subtle cues is particularly valuable in survey research, where responses are often informal, inconsistent, or off-the-cuff. While embeddings and LLMs use this architecture differently—one producing numeric vectors, the other generating structured outputs—they both benefit from the deep contextual understanding transformers make possible. That shared foundation is what allows these models to transform open-end analysis from a manual bottleneck into a scalable, AI-assisted workflow.

EMBEDDINGS MODELS

Embeddings models are a class of transformer tools that translate text into a vector of numbers that represent its underlying semantic meaning. In this way, they are conceptually similar to principal component analysis (PCA), which reduces complex data into latent components that capture its underlying structure. The resulting vectors, often spanning hundreds of dimensions, position pieces of text in a high-dimensional space, where semantic similarity is reflected by the directional closeness between vectors, typically measured using cosine similarity.

Such a model yields something quantitative for the researcher to leverage from otherwise unstructured data. It allows open-ended responses to be compared not just by shared words or phrases, but by shared meaning. These vector embeddings can then be passed to other machine learning methods to map and summarize relationships between responses. Two people may use completely different language to express similar ideas. One might mention "faster checkout," while another says "don't want to wait in line." The wording differs, but both reflect the same core desire for speed and convenience. This relationship will likely be captured in the mathematical similarity between their embedding vectors.

There are many embeddings models available to researchers today. Some, like OpenAI's text-embedding-3-large or Cohere's embed-multilingual-v3, are hosted models accessed through APIs. Others, such as BERT, RoBERTa, and models from the SentenceTransformers library, are open-source and can be run locally, offering greater control over customization, transparency, and data privacy. These models differ in how they were trained, how much context they can incorporate, and how they structure their output. Which model performs best often depends on the task and application context.

At the time of this writing, we prefer OpenAI's text-embedding-3-large, which performs well across a variety of use cases and is relatively easy to implement. A useful feature of this model is that it allows the user to specify the number of output dimensions. This flexibility helps researchers avoid the curse of dimensionality, where extremely high-dimensional data can dilute meaningful patterns and degrade clustering performance. Reducing the dimensionality of embeddings can also improve computational efficiency and make results easier to interpret, particularly when visualizing relationships between responses. For models that don't offer configurable output size and use too many dimensions, we recommend using a technique like UMAP (Uniform Manifold Approximation and Projection) to reduce dimensionality of the output while preserving the structure of the semantic space.

AI models are evolving quickly. We recommend that researchers periodically consult the Massive Text Embedding Benchmark (MTEB; Muennighoff et al., 2023) which provides a regularly updated comparison of model performance across tasks like clustering, classification, and semantic search. It's a helpful resource for evaluating tradeoffs between different options and staying informed about newly released models.

While embeddings can be powerful even with minimal text, the quality of input still matters. For short survey responses, adding just a bit of light contextual framing, such as prefixing with the context of the response, "Customer Suggestion:" before embedding can help the model interpret meaning more effectively. Too little context can leave the model guessing; too much can dilute the semantic signal.

To cluster the embeddings, we use hierarchical clustering on a cosine dissimilarity matrix. Because hierarchical clustering is a distance-based algorithm, we first need to map the distances between vectors. In this context, cosine similarity is preferred over Euclidean distance because it captures the *direction* of the vectors rather than their *magnitude*, making it more effective for comparing the meaning of text regardless of response length or word count. While Euclidean distance measures the straight-line distance between two points, cosine similarity measures how aligned two vectors are in the embedding space.

For two embedding vectors **A** and **B**, cosine similarity is defined as

$$\text{cosine similarity} = \frac{A \cdot B}{|A||B|}$$

The result ranges from -1 to 1, where 1 indicates perfect alignment (maximum similarity), 0 indicates no similarity (orthogonal vectors), and -1 indicates opposite directions. In most natural language applications, embedding vectors are non-negative and semantically aligned, so similarity scores typically fall between 0 and 1.

In practice, distance-based clustering algorithms require a dissimilarity measure so that it behaves like Euclidean distance, so we convert cosine similarity into cosine dissimilarity:

$$\text{cosine dissimilarity} = 1 - \frac{A \cdot B}{|A||B|}$$

This transformation produces a metric where higher values indicate greater semantic distance, making it compatible with clustering methods like hierarchical clustering.

Embeddings, when paired with clustering methods, are especially effective at identifying *identity groups*, clusters of responses that express the same idea using different language. This is particularly useful in tasks like unaided brand recall, where respondents might refer to the same brand in multiple ways (e.g., “Coke,” “Coca-Cola,” “Diet Coke,” “Coca-Cola Zero”) but still mean the same thing. However, they tend to struggle with *abstract or conceptual groupings*. For instance, semantically related responses might be clustered into narrowly defined groups without recognizing that they all contribute to a broader theme. A model might correctly separate references to “speed,” “efficiency,” and “ease of use” into distinct groups, without recognizing that these ideas could all fall under a higher-level concept like “convenience.” This limitation arises because embeddings are trained to *capture local semantic similarity*. Off-the-shelf models, with no fine tuning, do not explicitly learn or represent hierarchical or categorical relationships. Without additional modeling or human interpretation, embeddings alone are unlikely to infer taxonomy structure or organize responses at a more conceptual level.

Large Language Models

While embeddings are effective at grouping responses based on surface-level similarity, large language models (LLMs) offer a more flexible and powerful tool for interpreting open-ended survey data. LLMs are generative AI models trained to predict the next word in a sequence. This simple-sounding objective, when scaled across massive datasets and billions of parameters, gives these models the ability to understand language, follow instructions, generate structured outputs, and reason across multiple levels of abstraction.

Unlike embeddings models, which encode text into static numerical vectors, LLMs operate through prompt-response interactions. When given a carefully constructed prompt (often including instructions, examples, and unstructured text), they can return summaries, classifications, structured taxonomies, or explanations in natural language. This makes them particularly well suited for workflows that require interpretation, abstraction, or the creation of human-readable themes.

One practical consideration when working with LLMs is batching. For longer lists of open-ended responses or multi-step workflows, it's often necessary to divide data into smaller batches and send them through the model in sequence. While it can be tempting to fit as much data as possible into a single prompt, performance often degrades when the input becomes too long, even within the allowable token limit. Models may lose focus, overlook examples, or return inconsistent results when overloaded. Larger batches can help the model generalize, but they must be sized carefully to balance context with reliability.

Another key advantage of LLMs is their ability to work with structured input and output formats, such as JSON. By clearly instructing the model to return results in a machine-readable structure, researchers can improve consistency, reduce ambiguity, and make the outputs easier to use in downstream pipelines. For example, a prompt might instruct the model to classify a response and return both a "label" and a short "justification" in JSON format, something that can be easily parsed, validated, or appended to an existing dataset.

Effective use of LLMs depends heavily on prompt engineering. A good prompt gives the model clear, concise instructions, includes examples when needed, and avoids unnecessary complexity. Adding too many rules or edge cases can confuse the model, while overly vague prompts may lead to inconsistent or generic output. In our experience, simpler prompts with well-defined formatting instructions tend to work best. Asking the model to return structured fields, keeping language concise, and validating edge cases iteratively are all part of the design process. Prompt engineering is less about writing perfect instructions up front and more about refining them over time based on how the model behaves.

LLMs are especially valuable when the task requires abstraction, labeling, or interpretation; tasks that go beyond recognizing semantic similarity. They can generate meaningful themes from scratch, organize responses into taxonomies, and classify responses even when the connection to a theme is implicit. When guided by clear instructions and evaluated with human oversight, LLMs can serve as a highly flexible tool for making sense of open-ended responses at scale.

Methodology Overview

The purpose of this analysis was to evaluate how well transformer-based models can code open-ended survey responses with accuracy, consistency, and practical usability. We focused on two AI-driven workflows: one primarily using embeddings for semantic clustering, and one using a large language model (LLM) for direct classification and theme generation. However, both workflows leverage LLMs to some degree.

Our goal was to understand how these tools perform in a real-world setting where open-ended responses are short, unstructured, and often vary widely in quality and clarity. We wanted to test whether these models could generate meaningful, interpretable themes at a level that would be acceptable (if not preferable) for use in survey research and reporting.

To explore this, we applied both workflows to a large set of open-ended responses drawn from an applied market research study. The next section provides an overview of the dataset, including the question asked, the volume of responses, and the kinds of variability observed in the raw text.

The sections that follow then describe the two modeling workflows we tested: first, a bottom-up clustering approach using embeddings; and second, a top-down reasoning approach using a large language model.

About the Data

The dataset used in this analysis comes from a syndicated market research study conducted by The Harris Poll, focused on emerging automotive technologies. The study surveyed over 10,000 U.S. vehicle owners and included a variety of closed- and open-ended questions. For the purposes of this paper, we focused on a single open-ended item that asked:

> "What vehicle features or technologies, if any, do you feel you want or need that your current vehicle is not equipped with? These could be existing technologies you are aware of, or something not yet on the market. Please be as descriptive as possible."

This question yielded a large and diverse set of responses. Some were short and vague (e.g., "none," "not sure"), while others provided detailed lists of multiple technologies. Many responses exhibited inconsistent grammar, spelling, or punctuation, and varied widely in clarity and specificity. Some respondents described one clear feature; others mentioned several unrelated ideas. In total, the dataset contained thousands of open-ended responses suitable for analysis.

We intentionally selected this question because it reflects the real-world challenges of open-end analysis: short, variable-length responses, inconsistent formatting, and a lack of standardized terminology. These characteristics make it a strong test case for evaluating how well transformer-based models can identify themes, group similar responses, and produce useful outputs at scale.

Feature Extraction and Input Preparation

Before applying either the embeddings or LLM-based workflow, we performed a focused feature extraction step to clarify what exactly each model should be attending to within each open-ended response. This step was essential for reducing noise, improving interpretability, and ensuring that downstream models would focus on the most relevant content.

Although open-ended responses often contain multiple ideas, filler language, or contextual details that may not be essential for analysis, we found that extracting a concise phrase or set of keywords from each response led to better performance in both modeling pipelines. These extracted features acted as focused signals that highlighted the core of what the respondent was trying to communicate.

The feature extraction process was partially automated using an LLM with prompt-based instruction. For each raw open-end, we asked the model to extract the key technologies, features, or desires being expressed.

Responses such as:

> "I guess something like better safety would be nice, and maybe more cameras, but I'm not really sure."

were converted into simpler, focused phrases like:

> "Safety features, additional cameras"

This type of transformation allowed both downstream workflows to work from a cleaner and more consistent signal. In practice, we found that this step not only improved the quality of clustering and classification but also made human review easier by providing a structured starting point.

In both the embeddings and LLM workflows, this feature-level input became the core unit of analysis. Whether grouping semantically similar ideas or classifying them into higher-level categories, the workflows benefited from starting with a distilled representation of the respondent's original answer. While helpful across both approaches, this focused input was especially important for the embeddings workflow, where the quality of clustering depends heavily on both the clarity and relevance of the input. Without this step, vague, multi-topic, or off-focus responses introduced noise that diluted the structure of the output. Feature extraction ensured that only the specific vehicle technologies or features being mentioned were used as input, which improved the consistency and interpretability of the resulting clusters.

In addition to simplifying and clarifying the language, the feature extraction step also served to separate multi-topic responses into individual features. Many open-ended responses contained references to more than one idea, for example, a respondent might mention a desire for both safety improvements and entertainment upgrades. Rather than embedding or classifying the full response as a single unit, we extracted each meaningful concept as its own feature. This allowed downstream models (particularly in the embeddings workflow) to group semantically similar features more accurately. It also had the practical effect of approximating fuzzy clustering, where one respondent's ideas could contribute to multiple thematic groups, even though each extracted feature was treated as a separate input.

Altogether, the result was a cleaner, more focused dataset that improved the quality of downstream analysis in both workflows. The extracted features became the core unit of analysis, offering a more structured, simple, and interpretable foundation than the original free-text responses.

EMBEDDINGS WORKFLOW

The embeddings workflow used a bottom-up clustering approach to group features based on semantic similarity. Each extracted feature was converted into a numeric vector using a transformer-based embedding model. To help the model interpret the input consistently, each feature was prefixed with a short label (e.g., "Vehicle Feature:") to provide minimal context.

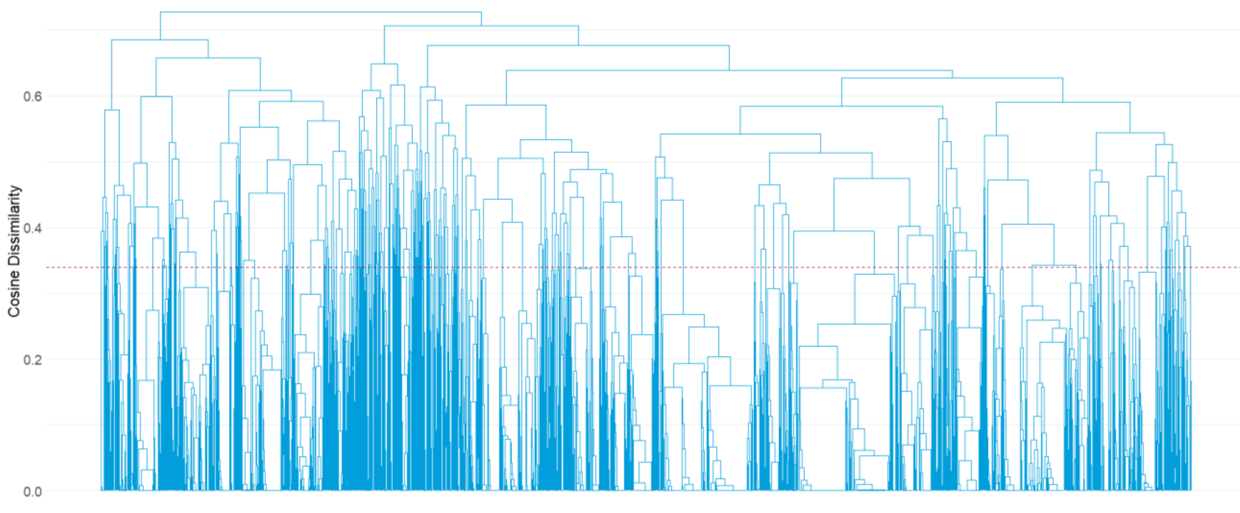
We calculated pairwise cosine dissimilarities between features to form a distance matrix. Cosine dissimilarity was used instead of Euclidean distance because it reflects directional similarity, which is more aligned with semantic meaning.

We then applied hierarchical clustering to the dissimilarity matrix. This method does not require a fixed number of clusters and allows for flexible control over granularity. We selected a distance threshold ($h = 0.34$) based on iterative review, adjusting the cutoff until the output grouped responses in a way that matched human expectations. Clusters were manually labeled using a sample of features from each group.

Once clustering was complete, we used a large language model to assist in labeling each group. The LLM was prompted with a sample of features from each cluster and asked to suggest a concise label. These suggestions were reviewed and edited by a human to ensure clarity and consistency. The final output was a binary-coded dataset showing which features belonged to which themes.

Figure 1 below shows the histogram of the cosine dissimilarity of embeddings.

Figure 1



LLM Workflow

The second approach used a large language model (LLM) to directly classify and organize extracted features into thematic groups. Rather than relying on vector-based clustering, this workflow treated the task as a prompt-driven classification and labeling problem.

We began by designing a prompt that instructed the LLM to assign each feature to one or more high-level themes based on its content. To create the thematic structure, we first processed a large batch of features through the model and asked it to generate a list of common categories observed in the data. These categories served as the starting point for the coding frame. Once the categories were established, we used the same model to classify each feature into one or more of the identified themes.

Due to the limited context window of the model, we batched the classification process into manageable segments. Features were grouped into batches of 100–150 inputs and processed sequentially. We observed that model accuracy degraded when input size approached the context limit, even when technically allowed, so we favored smaller batches with clearer context.

To improve consistency and enable automation, we instructed the model to return results in structured JSON format, including the assigned theme and a short justification for each classification. This made the output easier to parse, review, and integrate into downstream processes.

Prompt design was iterative. We found that concise instructions, consistent formatting, and light examples led to better results than overly detailed rules. After initial testing, we locked in a standardized prompt template that balanced clarity with model flexibility.

Final classifications were reviewed by a human for quality assurance. While most assignments aligned well with expectations, some themes required relabeling, merging, or refinement to eliminate redundancy or ambiguity.

Results

Both workflows found similar top features. See Figures 2 and 3 for the most commonly mentioned vehicle technologies.

Figure 2

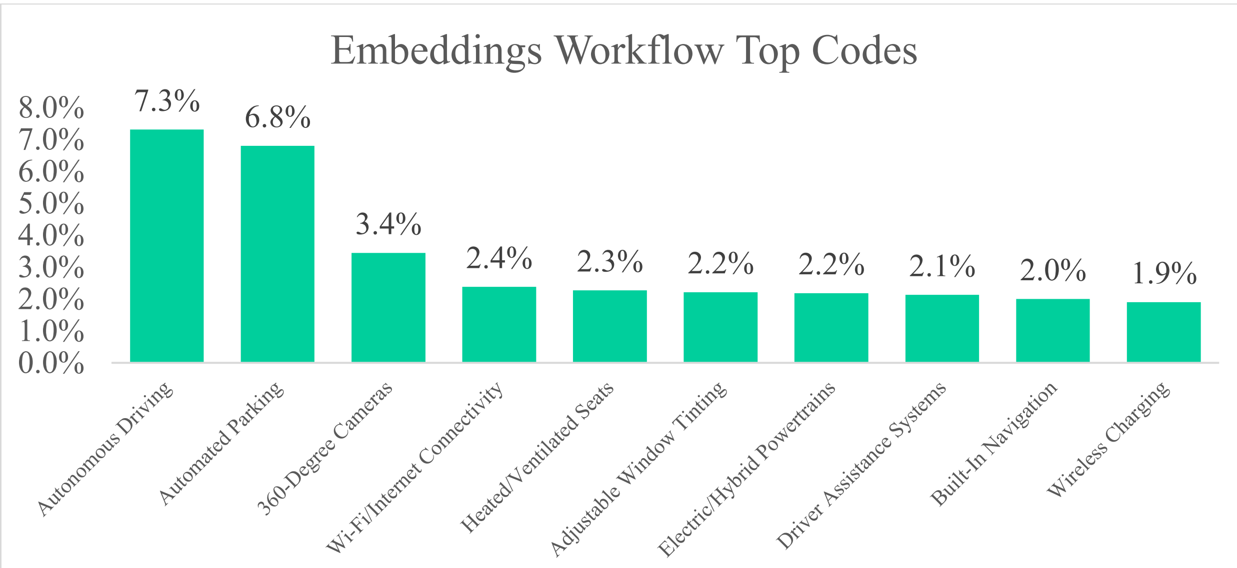
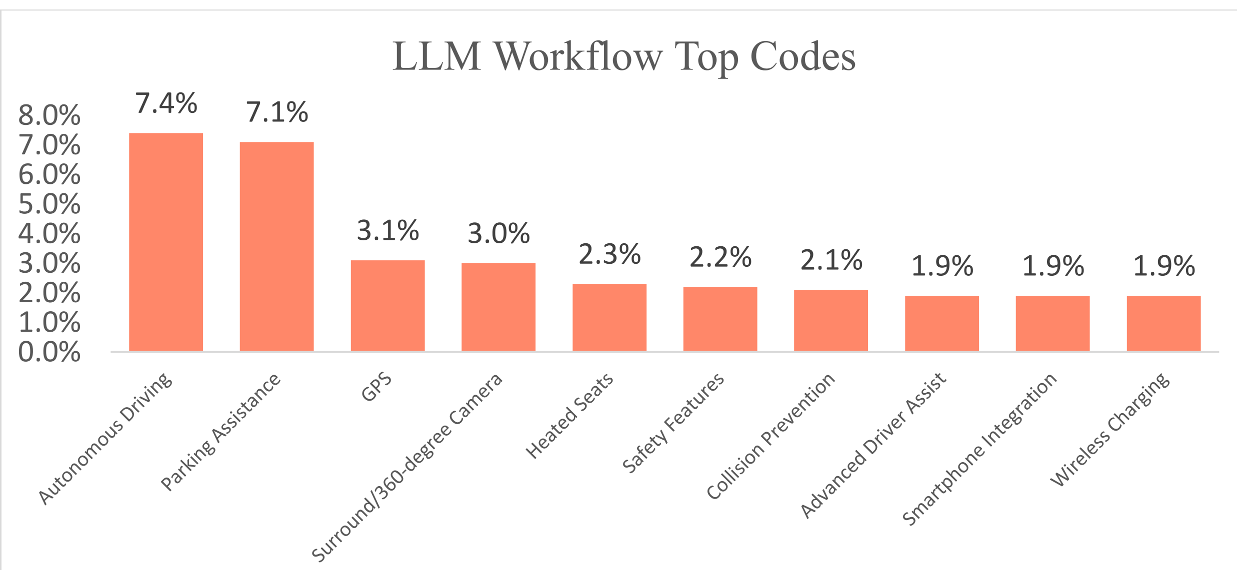


Figure 3



One notable finding: nearly 40% of responses were effectively “None” (e.g., “No new technology desired”), highlighting the importance of distinguishing between meaningful feedback and non-substantive answers in large-scale text data.

While the top codes were very similar, each method identified different numbers of codes. For the embeddings workflow we identified 224 codes, while the LLM workflow identified 129 codes. Both codebooks generated had their merits, with the embeddings workflow being much more specific, and the LLM workflow codes allowed to be more abstract and conceptual, and preference for either would depend upon the researcher’s goals.

While the high-level codes were aligned, the workflows produced codebooks of very different granularity. The embeddings workflow surfaced 224 distinct codes, while the LLM workflow yielded 129. The embeddings approach produced more specific, tightly grouped codes, while the LLM-generated codes leaned more abstract and thematic. Which approach is preferable depends on the goals of the analysis . . . whether the focus is on detailed feature-level differentiation or broader conceptual themes.

To assess coding quality, we conducted a validation exercise using a human coder. The coder was given both AI-generated codebooks and asked to manually tag a sample of the original responses. We then compared these human-coded results against the AI assignments. Table 1 summarizes the outcome:

Table 1

Comparison	Embeddings Workflow	LLM Workflow
Perfect Agreement	81%	75%
Close Agreement / Ambiguous	6%	8%
Human Coding Preferred	7%	9%
AI Coding Preferred	5%	9%

These results suggest that both methods performed well overall, with minor tradeoffs in accuracy versus abstraction. Importantly, most discrepancies were not clear “errors” but stemmed from nuanced differences in interpretation or codebook structure.

In several cases, disagreements came down to having no ideal label rather than a true misclassification. As in most coding tasks, parsimony must be balanced with comprehensiveness, and simplification often leads to some loss of nuance. Still, we noted occasional gross mismatches that, with further refinement, could likely have been corrected via better prompts or taxonomy design.

Where the methods differed most was in *why* they disagreed. The embeddings workflow tended toward over-specification. Its codebook was more rigid, with fine-grained distinctions like “GPS” vs. “Navigation System” or “Bluetooth” vs. “Smartphone Integration.” While technically separate features, these concepts often overlapped in how respondents described them. This specificity introduced ambiguity during manual review.

The LLM workflow had the opposite challenge. It favored broader categories, which sometimes led to overlapping or vague themes. This generality made it harder to maintain a strict taxonomy, especially when a single feature could belong to multiple categories.

Overall, the embeddings model was better at grouping responses that expressed similar ideas using different language (identity resolution), while the LLM excelled at organizing responses into conceptual hierarchies. A hybrid approach, where embeddings are used to cluster identity groups and LLMs are used to label and abstract them, was untested at the time of this conference, but may offer the best of both worlds.

Conclusion

AI is ready to help researchers code open-ends but only when used thoughtfully and with strong human supervision.

Transformer-based models like embeddings and large language models (LLMs) offer powerful new ways to scale the analysis of open-ended survey responses. Each method has its strengths. Embeddings workflows tend to be more structured and specific, making them effective for grouping similar responses at a fine-grained level. But this same rigidity can introduce ambiguity and create distinctions that may not be meaningful in practice. LLM workflows, on the other hand, offer flexibility and abstraction, allowing for more conceptual coding, but they can struggle to maintain a consistent taxonomy and often produce overlapping or vague themes.

When we tested both approaches on real-world survey data, we found that:

- Both workflows surfaced similar top codes.
- Embeddings produced more codes overall, but with narrower distinctions.
- LLMs grouped responses into broader, higher-level themes.
- Both approaches reached 75–81% perfect agreement with human coding on first pass, on large code books.

Most disagreements weren't true errors, they were the kinds of judgment calls researchers make all the time when interpreting raw data. And in many cases, we found that the best results came from combining the strengths of both models: using embeddings to cluster semantically similar responses, and LLMs to label or abstract those clusters into meaningful, reportable themes.

Importantly, this isn't just experimental anymore. Commercially available tools are already emerging that refine these workflows and provide user-friendly platforms for working with AI and open-ends. These solutions make it easier than ever to integrate AI into everyday research processes, allowing researchers to spend less time cleaning data and more time understanding it.

Ultimately, AI won't replace human coders, but it will change their role. Instead of reading and tagging every response by hand, researchers become editors, validators, and strategic guides. AI does the heavy lifting; humans focus on judgment, refinement, and context.

That shift unlocks significant value. Researchers can include more open-ends without adding weeks to their timeline or straining their budgets. They can uncover themes faster, spot emerging topics earlier, and focus their attention on interpreting the story behind the data, not wrangling it.

And this is just the beginning. As models improve and interfaces evolve, we expect workflows to become more seamless, with better tools for interactive review, domain-tuned and other fine-tuned models, and active learning systems that refine themselves over time. Open-end analysis is being transformed . . . and AI is no longer a moonshot. It's a working tool, ready to use today.

Jacob Nelson

References

Adrogue, F. (n.d.). Clustering open-ended questions: The algorithm to automatically quantify speech. In Analytics & Insights Summit (pp. 111–120). Barcelona, Spain.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022. https://doi.org/10.7551/mitpress/1120.003.0082

Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2023). MTEB: Massive text embedding benchmark. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (pp. 2014–2037). Dubrovnik, Croatia: Association for Computational Linguistics.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3982–3992). Hong Kong, China: Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, 30.

DEEPENING CONSUMER INSIGHTS—CONVERSATIONAL AI

MOHIT SHANT
RAJAT NARANG
MD. FAISAL
INSIGHTS CURRY

ABSTRACT

Conversational Artificial Intelligence (AI) is transforming the way market researchers engage with respondents, replacing static, one-directional survey formats with dynamic, adaptive dialogues. This paper explores how conversational AI can enhance engagement, trust, and decision-making in high-stakes consumer contexts, drawing on a live deployment in the real estate sector. Using Agent Mira—a context-aware, multi-agent conversational AI integrated into a real estate app—we conducted a proof-of-concept study with 154 active homebuyers. The system guided participants through property evaluations, clarified questions, and provided personalized insights such as tailored offer prices and win probabilities. Results showed exceptionally high engagement (99%), substantial influence on decision-making (81%), and a notable capacity to convert AI skeptics into advocates (83% expressed willingness to use AI in future interactions). Building on both academic literature and our case study, we identify best practices for designing conversational AI systems for market research, discuss ethical guardrails, and outline cross-industry applications. The findings indicate that conversational AI has the potential not only to improve response quality and data richness but also to humanize digital interactions at scale.

1. INTRODUCTION

The practice of market research has long depended on structured instruments such as questionnaires and surveys to gather consumer insights. While effective in many respects, these tools are often limited by their static nature. Respondents may face ambiguity in question wording, lack opportunities to seek clarification, and may disengage if the process feels mechanical or irrelevant. In parallel, response quality has faced challenges due to issues such as inattentive respondents, "straight-lining," and the proliferation of bots in online surveys.

In the past decade, advances in conversational AI—systems designed to simulate human-like interactions through text or speech—have offered an alternative approach. By enabling two-way, adaptive dialogue, conversational AI can replicate many qualities of human interviewers: clarifying questions, adapting to context, and responding with empathy. Applications in customer support, healthcare, e-commerce, and human resources have already demonstrated that AI-driven conversations can boost engagement, deliver personalized experiences, and build trust.

This paper focuses on the application of conversational AI in market research, specifically in a high-involvement decision-making context: purchasing a home. Home buying involves complex information processing, emotional considerations, and significant financial stakes. This makes it a strong testing ground for whether AI can truly enhance decision-making and humanize digital research interactions.

We introduce **Agent Mira**; a conversational AI system embedded in a real estate app, designed to guide users through property evaluation, mortgage considerations, and neighborhood assessments. Our objective is to demonstrate, through a combination of literature grounding and empirical results, that conversational AI is not only a technical innovation but also a methodological shift in marketing research.

2. Background and Context

2.1 Conversational AI in Consumer Interaction

Conversational AI refers to artificial intelligence systems capable of natural, context-aware, and adaptive dialogue with humans. This category includes chatbots, voice assistants, virtual agents, and generative AI-based interfaces. Research has shown that conversational systems can improve user satisfaction by responding with empathy, providing personalized recommendations, and adapting to user needs in real time (Shawar and Atwell, 2007; McTear, 2020).

In commercial contexts, customer support bots reduce service costs while improving response speed. E-commerce platforms deploy conversational agents to guide product discovery and purchase decisions (Gnewuch et al., 2017). In healthcare, AI chatbots assist with symptom checking and patient triage (Bickmore et al., 2010). These successes suggest that similar principles could be applied to market research, where engagement and clarity are critical.

2.2 Limitations of Traditional Surveys

Traditional surveys are typically static, one-directional, and constrained by predetermined question sequences. Respondents cannot easily seek clarification, and ambiguous wording may result in inaccurate responses. Engagement is often limited, with respondents providing minimal effort answers or abandoning surveys entirely (Revilla and Ochoa, 2017). Furthermore, the rise of automated bot responses has degraded data quality in online survey platforms (Kennedy et al., 2020).

2.3 Conversational Surveys: A New Paradigm

Conversational surveys replace rigid questionnaires with dynamic dialogues. The AI adapts questions based on prior responses, offers clarifications, and can detect inconsistencies or low-effort responses in real time. The interaction becomes a two-way exchange rather than a unidirectional extraction of data. Prior studies suggest that conversational surveys can increase response completeness, data richness, and participant satisfaction (Cavallini et al., 2021).

2.4 Trust and Decision Influence

Trust is a critical factor in AI adoption. In decision-making contexts, users are more likely to rely on AI outputs when they perceive them as accurate, transparent, and aligned with their needs (Dietvorst et al., 2015). Research also shows that follow-up questions can signal attentiveness and personalization, further strengthening trust (Glikson and Woolley, 2020). In our case study, trust dynamics are central: homebuyers' willingness to follow AI guidance in a high-stakes purchase provides a robust test of AI's influence.

3. METHODOLOGY

3.1 System Overview: Agent Mira

Agent Mira is a multi-agent conversational AI system embedded within a real estate app. Unlike generic chatbots, it uses specialized agents for distinct tasks—such as pricing, terminology clarification, and market trend analysis—coordinated through a framework ensuring consistency. The system is context-aware, meaning it adapts its dialogue based on the user's prior inputs and stage in the homebuying journey.

3.2 Key Design Principles

- **Context Awareness:** Builds on each respondent's prior answers to maintain conversational flow, adapt subsequent questions, and minimize redundancy. This continuity helps improve engagement and data accuracy.
- **Task Specialization:** Deploys dedicated agents for different areas of expertise (e.g., compliance, data validation, clarification), ensuring deeper insights and more precise responses.
- **Ethical Guardrails:** Operates within clearly defined boundaries to prevent speculative, biased, or non-compliant content, aligning with privacy regulations and ethical standards.
- **Iterative Learning:** Continuously refines question phrasing, engagement strategies, and decision logic through testing, feedback loops, and performance monitoring.
- **Seamless Integration:** Embeds the conversational mechanism directly into the platform's interface, allowing users to interact without leaving their workflow and supporting consistent branding.

3.3 Proof-of-Concept Survey

A proof-of-concept (POC) survey was conducted among 154 respondents actively seeking to purchase a new home. The demographic profile included:

- **Age:** 24–56 years
- **Gender:** 70% male, 30% female
- **Income:** Primarily $125,000+ annual income
- **Tech Comfort:** Varied, from minimal to high

The survey process involved:

1. **Property Evaluation**—Users explored listings and discussed amenities.
2. **Personal Preferences**—Respondents shared requirements, budget constraints, and location preferences.
3. **Offer Price Calculation**—Agent Mira provided personalized offer price suggestions, win probability estimates, and rationale.

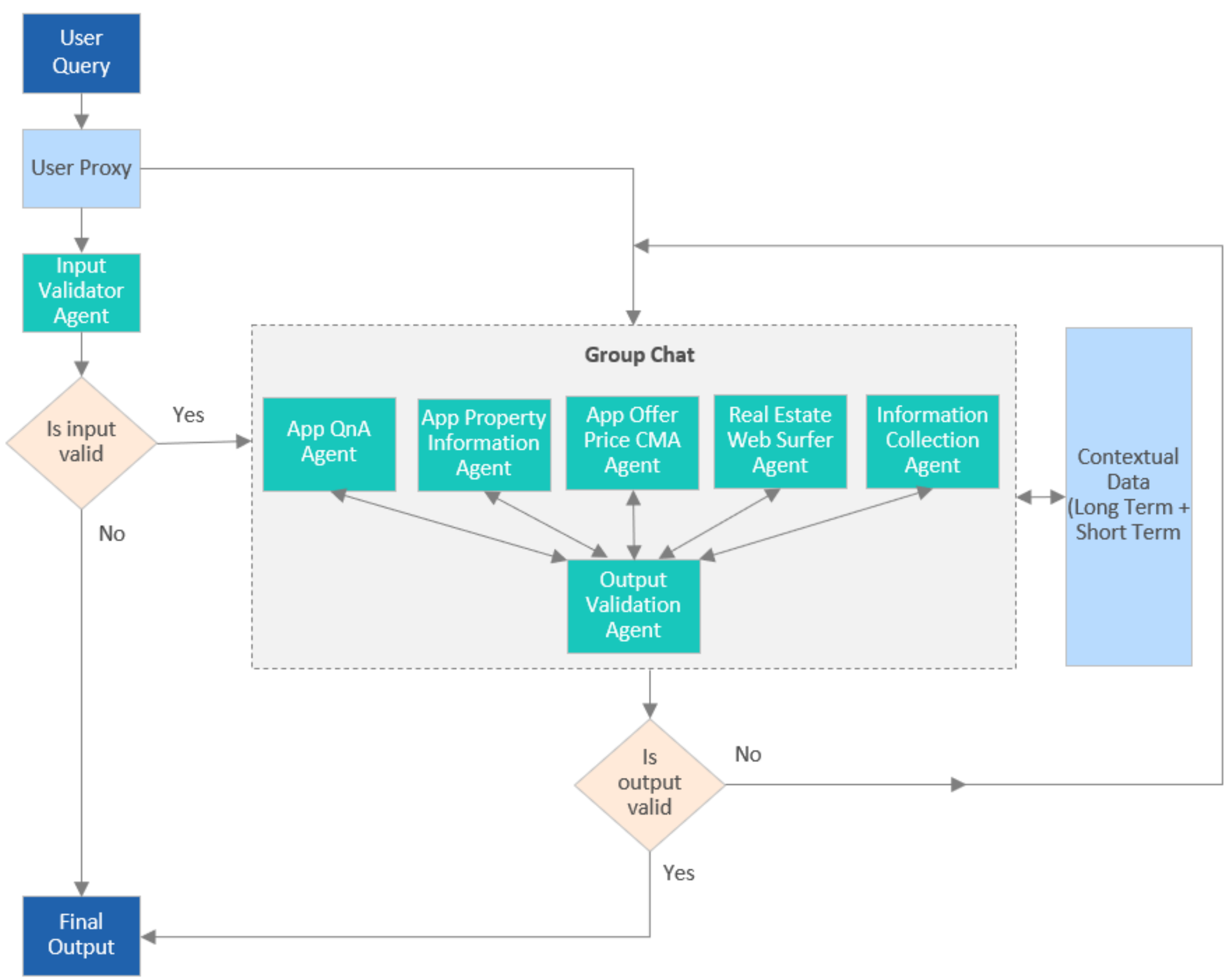
Data was collected on:

- Interaction rates
- Perceived helpfulness
- Trust levels
- Influence on decision-making

3.4 Architecture and Agent Roles

To operationalize the proof-of-concept survey, Agent Mira was deployed through a multi-agent workflow designed to ensure input validity, contextual alignment, and compliance before final output delivery. As illustrated in Figure 1, user queries were first screened through an input validation layer and then processed collaboratively by specialized agents within a group chat environment. Each agent contributed domain-specific expertise, while a shared contextual memory supported continuity and personalization. The workflow concluded with output validation, ensuring that responses met both quality and regulatory requirements.

Figure 1



Agent Definitions

The system is composed of specialized agents, each designed to fulfill a distinct role:

- **Validator Agent**—Validates user inputs for compliance, screens for personally identifiable information (PII), toxicity, and inappropriate content.
- **App QnA Agent**—Clarifies app-related questions, guides navigation, and explains usage-related queries.
- **Real Estate Agent**—Provides domain knowledge by sharing insights on property risks, market trends, and general real estate queries.
- **Information Collection Agent**—Captures user-provided information such as preferences, constraints, and contextual details, storing them for short- and long-term use.
- **Offer Price and Property Information Agent**—Specializes in property valuation and offer price recommendations, integrating comparable data and market indicators.
- **Output Validation Agent**—Validates all responses for compliance with Fair Housing standards, responsible AI practices, and appropriateness before presentation to the user.

Integration of Contextual Data

All agents are supported by a shared contextual memory that stores both long-term and short-term user data. This enables continuity across interactions, reduces redundancy in questioning, and strengthens personalization in recommendations.

Together, this architecture transforms the system from a simple chatbot into a coordinated team of specialized agents, ensuring rigor, compliance, and decision support in real estate interactions.

4. Results

4.1 Engagement and Interaction Rates

The proof-of-concept (POC) study recorded a **99% interaction rate** with Agent Mira, meaning nearly every respondent engaged with the AI during their decision-making process. This is a substantial improvement over traditional online surveys, where completion rates often fall below 50% (Kaplowitz et al., 2004).

While the study design does not allow us to determine causality, several factors may help explain this elevated engagement:

- Seamless Interface Integration—Because the conversational mechanism was embedded directly into the property browsing experience, respondents did not need to switch platforms.
- Contextually Relevant Dialogue—Tailoring questions to each user’s stage in the homebuying journey may have made the interaction feel more timely and personal.
- Interactive Decision Support—Providing actionable insights such as offer price suggestions and market comparisons likely encouraged participation beyond data entry.

Taken together, these patterns suggest that integrating conversational AI into a user's natural workflow *can be associated with* improved engagement, but additional controlled studies would be required to isolate the precise drivers.

4.2 Influence on Decision-Making

A key objective of the study was to assess whether AI could meaningfully shape high-stakes consumer decisions in real estate. Results indicate that **81% of respondents** reported the AI's insights had a significant influence on their choices, most notably in offer pricing and neighborhood selection.

Trust levels varied depending on the buyer's stage:

- **Late-stage buyers** (ready to submit an offer) reported the highest confidence, with **55–75% rating the AI as "extremely helpful."**
- **Early-stage buyers**, who were still exploring options, reported lower reliance, with **25–26% rating the AI as "extremely helpful."**

One possible interpretation is that later-stage buyers had more domain knowledge, making it easier to verify the AI's suggestions, which in turn may have increased trust. Alternatively, their higher decision urgency could have amplified their receptivity to guidance. These findings indicate that conversational AI has the potential to influence decision-making, though the mechanisms underlying this influence warrant further exploration.

4.3 Trust Dynamics

Trust in AI is frequently cited as a barrier to adoption, particularly in high-stakes domains like real estate. In this study, respondents who began with skepticism demonstrated notable shifts after interacting with Agent Mira:

- **78%** rated the experience as *"more engaging"* than expected.
- **83%** expressed they would *"definitely"* want conversational AI in future interactions.
- **72%** reported the AI significantly influenced their decisions.
- Only **2%** maintained that they would avoid AI in the future.

These results suggest that positive, personalized experiences can *contribute* to trust formation. While the study cannot confirm causality, the findings align with prior research on *experiential trust development* (Glikson and Woolley, 2020), which highlights that hands-on exposure can rapidly shift perceptions when combined with transparency and value delivery.

4.4 Perceived Helpfulness and Follow-Up Questions

Follow-up questioning emerged as a key factor linked to engagement and perceived value:

- **88%** of respondents felt that follow-up questions improved their experience.
- Among initially distrustful participants, **67%** rated the AI as *"extremely helpful"* after receiving follow-up prompts (compared to **37% overall**).

Although we cannot attribute causality, the AI's ability to clarify terminology, provide contextual insights, and adapt recommendations in real time likely contributed to these perceptions. These findings support the idea that attentive, iterative dialogue signals personalization and reliability, thereby enhancing user satisfaction and decision confidence.

4.5 Data Quality Improvements

Conversational surveys enable real-time checks that can improve data quality compared to static surveys. In this study, the AI:

- Flagged incomplete or inconsistent answers for clarification.
- Provided immediate feedback when random or contradictory responses were detected.
- Reduced the incidence of low-quality data entries relative to traditional survey benchmarks.

These results suggest that conversational formats may strengthen data integrity, though future work is needed to quantify the magnitude of these improvements across larger samples.

5. Discussion

5.1 From Static to Dynamic Research

The findings reinforce the literature's assertion that conversational surveys can address many of the inherent limitations of traditional instruments. Agent Mira's dynamic, context-aware dialogues improved engagement by maintaining relevance throughout the interaction. This interactivity encouraged respondents to elaborate on their answers, ask clarifying questions, and provide richer detail—resulting in higher-quality datasets that would be difficult to obtain through static, one-directional formats.

5.2 Trust as a Design Outcome

Trust was not an incidental byproduct of system performance—it was an intentional design goal. Beyond technical accuracy, trust was cultivated through contextual awareness, domain-specific task specialization, and clearly defined ethical guardrails. The system's use of empathy, transparency, and actionable recommendations was particularly effective in converting initially skeptical users. This aligns with Parasuraman et al.'s (2000) technology readiness framework, which emphasizes assurance and control as critical adoption drivers.

5.3 Influence on High-Involvement Decisions

The real estate context amplified the significance of these findings. High-involvement decisions typically involve extensive information search and evaluation. That respondents reported a direct influence of AI on their offer pricing strategies demonstrates that conversational AI can move beyond a data collection tool to become a trusted advisor.

5.4 Implications for Market Research

For the marketing research community, these findings highlight several actionable principles for designing and deploying AI-enabled research systems:

- **Engagement is not just a metric**—Sustained engagement directly drives data completeness and richness, reducing drop-off rates and improving the representativeness of the dataset.
- **Trust-building must be intentional—**Features such as transparency, contextual awareness, and clear rationale for outputs should be built into AI interactions from the outset, not treated as optional enhancements.
- **Adaptive questioning delivers dual benefits—**Dynamically adjusting question sequences based on prior responses both increases respondent satisfaction and improves the relevance and quality of collected data.
- **Context integration is critical—**Leveraging previous responses, behavioral signals, and situational factors ensures that the interaction feels personally relevant and produces more meaningful insights.

By adopting these principles, market researchers can move beyond static data collection toward richer, relationship-driven engagements that not only gather information but also influence consumer decision-making.

6. Best Practices for Designing Conversational AI in Research

The deployment of Agent Mira revealed that the difference between a merely functional chatbot and a transformative research tool lies in design discipline. Below, we detail six best practices, each grounded in both the literature and our empirical observations.

6.1 Ensure Contextual Awareness

A common frustration in static surveys is the lack of acknowledgement of prior answers. Agent Mira demonstrated that **remembering and referencing user inputs is central to building rapport and trust**. For example, when a respondent previously mentioned a preference for “quiet neighborhoods,” Mira later connected this to market insights by saying:

> “You mentioned wanting a quiet area—the last three sales here were in low-traffic streets, which could fit your preference.”

Literature Link: Context retention increases perceived intelligence and reduces cognitive load (Clark and Brennan, 1991).

6.2 Use Specialized Agents for Complex Domains

Rather than a monolithic AI, Mira used multiple specialized agents—pricing, terminology, trend analysis, validation—each optimized for a **specific expertise domain**. This modular approach ensured:

- Depth in answers.
- More precise data validation.
- Reduced risk of "hallucination" errors.

Implication for Researchers: A single generic model may suffice for simple customer feedback, but complex research requires **task-specific sub-agents**.

6.3 Implement Governance and Ethical Guardrails

AI without governance risks undermining trust. Mira's ethical framework ensured:

- No speculative investment advice.
- Compliance with housing regulations.
- Pre-programmed "safe exits" for sensitive questions.

Industry Takeaway: Governance should be a **co-equal design priority** alongside conversational flow.

6.4 Embed Continuous Learning Loops

Feedback mechanisms allowed respondents to rate AI helpfulness in real time. This served two functions:

- Immediate correction for the current conversation.
- Longer-term model improvement.

Observed Benefit: A respondent who initially rated a clarification as "unclear" was provided with a revised, simplified explanation—improving satisfaction instantly.

6.5 Integrate Seamlessly into Existing Interfaces

Agent Mira was **embedded natively** in the real estate app. Respondents never had to "switch context" or open a separate survey tool. This reduced friction and kept engagement high.

Practical Note: Embedding AI in the existing customer journey yields better results than bolting it on as a separate channel.

6.6 Detect and Manage Data Quality in Real Time

The Validation Agent flagged contradictory inputs (e.g., a $200k budget when all viewed properties exceeded $500k) and politely prompted clarification:

> "I noticed your budget is lower than the properties you've been exploring—should I adjust my search filters?"

This not only improved data accuracy but also enhanced perceived personalization.

7. Cross-Industry Potential

While the real estate sector provided an ideal stress test for Agent Mira, the **underlying principles of conversational AI in research are highly transferable**. Below are four sectors where such technology can be transformative.

7.1 Healthcare Research

Use Case: Patient-reported outcomes, pre-visit symptom capture, clinical trial feedback.
Conversational AI can:

- Translate medical jargon into plain language.
- Encourage more honest reporting due to reduced social desirability bias.
- Offer multilingual, culturally adapted questionnaires.

Example: A patient describing "feeling heavy in the chest" could be guided to clarify whether this is "shortness of breath" or "pressure"—improving diagnostic data quality.

7.2 Financial Services

Use Case: Understanding consumer attitudes toward savings, investments, and credit products.
Conversational AI can:

- Clarify complex terms ("APR," "compound interest").
- Assess risk tolerance interactively.
- Offer scenario-based questions for deeper insight.

Value: Financial literacy varies greatly; adaptive explanations ensure all respondents can engage meaningfully.

7.3 Product Development and Innovation

Use Case: Gathering consumer feedback on prototypes or new features.
Conversational AI can:

- Explore the "why" behind preferences.
- Conduct iterative A/B-style preference testing within the same conversation.
- Capture emotional reactions through sentiment analysis.

7.4 Government and Public Services

Use Case: Citizen engagement surveys, service feedback.
Conversational AI can:

- Increase accessibility for low-literacy populations.
- Provide multi-lingual options for inclusivity.
- Adapt tone and content to sensitive topics.

Impact: Higher-quality citizen input leads to better policy design.

8. Ethical Considerations

Deploying conversational AI in research raises a set of ethical challenges that must be addressed from the outset.

8.1 Transparency

Participants should be explicitly informed they are interacting with AI, and the AI should be capable of answering the question:

> "Are you a human or a bot?"

Failing to do so risks **erosion of trust** if participants discover the truth later.

8.2 Data Privacy and Minimization

Only collect what is necessary for the research purpose. Agent Mira, for example:

- Avoided storing sensitive personal identifiers unless explicitly provided.
- Purged conversation logs of identifying data before analysis.

Regulatory Note: Compliance with GDPR, CCPA, or local privacy laws is non-negotiable.

8.3 Bias Detection and Mitigation

Conversational AI models can inadvertently reflect societal biases in:

- Language tone
- Advice framing
- Interpretation of open-ended responses

Mitigation Strategy: Regular auditing of outputs for demographic disparities.

8.4 Informed Consent for Conversational Data

Unlike static survey answers, conversational transcripts contain **richer, more identifiable information**.
Consent forms should:

- Explain how data will be stored.
- Clarify who will have access.
- Specify if data will be used for AI training.

9. Future Directions

The evolution of conversational AI in research is still in early stages. Based on our findings, we foresee several key directions:

9.1 Multi-Lingual, Culturally Adaptive AI

Expanding to global markets requires:

- Accurate translation.
- Cultural adaptation of idioms, examples, and tone.
- Sensitivity to local norms in question sequencing.

9.2 Continuous-Learning AI Panels

Imagine an AI-maintained longitudinal research panel:

- The AI remembers prior responses.
- Follows up over time with context-aware questions.
- Adapts research themes based on individual and aggregate changes.

Benefit: Reduces panel fatigue and keeps engagement high.

9.3 Voice-Interactive Surveys

Adding voice capabilities:

- Increases accessibility for older adults or low-literacy users.
- Allows for emotion detection via vocal tone.
- Broadens device compatibility (smart speakers, IVR systems).

9.4 RealTime Sentiment Adaptation

Future conversational AIs could:

- Detect frustration or enthusiasm in real time.
- Adjust tone, pacing, or question framing accordingly.
- Provide researchers with richer emotional context alongside factual answers.

10. Conclusion

This study demonstrates that conversational AI can fundamentally transform marketing research by moving beyond static, one-way question-and-answer formats toward dynamic, adaptive, and trust-building dialogues. The deployment of Agent Mira in the real estate sector yielded **exceptionally high engagement rates**, richer datasets through elaboration and clarification, and measurable influence on high-stakes decision-making—even among participants who began with skepticism toward AI.

Key success drivers included **context awareness** to maintain continuity, **task specialization** through coordinated agents, **ethical guardrails** to ensure compliance, and **iterative learning loops** for continuous refinement. Together, these design elements created an experience that was not only functional but also **perceived as valuable and trustworthy**, converting hesitant users into advocates.

The implications extend well beyond real estate. Any industry facing the challenge of low survey completion rates, data quality concerns, or complex decision-support needs can adapt similar principles to drive deeper engagement and richer insights. By embedding conversational AI directly into the user's natural workflow, researchers can create **relationship-driven data** collection environments that generate higher-quality responses, improve trust in AI-assisted recommendations, and ultimately lead to more informed strategic decisions.

Looking ahead, the convergence of **multi-agent architectures**, **domain-specific tuning**, and **cross-platform integration** offers a pathway for scalable, human-like AI research assistants. Organizations that invest early in this approach will be positioned to capture richer, more actionable insights, foster stronger consumer relationships, and differentiate themselves in increasingly competitive markets.

Mohit Shant

Rajat Narang

Md. Faisal

REVISITING THE NO-CHOICE OPTION IN CONJOINT ANALYSIS

CHENG-YU HUNG
OHIO STATE UNIVERSITY
PETER KURZ
BMS MARKETING RESEARCH + STRATEGY
ROGER A. BAILEY
OHIO STATE UNIVERSITY
JOEL HUBER
DUKE UNIVERSITY
GREG M. ALLENBY
OHIO STATE UNIVERSITY

ABSTRACT

The validity of using conjoint analysis to conduct an economic evaluation of product characteristics rests on the inclusion of brand names, prices, and an outside "no-choice" option in the choice task. The no-choice option is included in case respondents determine that some other offering, not included in the conjoint choice task, is preferred to those that are included and that it would be better to hold on to their money and not make a purchase at that time. Selecting the no-choice option assumes that respondents have some level of knowledge of the value and prices of goods in the market. In this paper, we show that survey respondents may lack this information and make inferences about market prices from the conjoint exercise itself. We find that the lack of knowledge of competitive offerings and prices affects estimates of brand values but not the value of other product features. In addition, we discuss aspects of how a well-designed conjoint study mitigates the effects of this type of learning in conjoint analysis.

1. INTRODUCTION

Conjoint analysis in marketing takes an economic view of choice in which respondents can recall and construct their preferences for hypothetical offerings (Ben-Akiva et al., 2019; Manski et al., 1981). The minimal requirements for conducting a conjoint study to study marketplace demand is the inclusion of product brand and price, and an outside "no-choice" option. Brand is needed for respondents to imagine the offering and recall associated attributes and levels of performance. The brand name serves as a proxy for the unmentioned attributes of a brand in a conjoint study, and consumer knowledge of these features is what gives the brand names its value. Prices provide a metric to reconcile utility estimates to a common scale and are needed to compute measures of economic value, such as willingness to pay. However, conjoint studies usually include other product attributes or features of interest to the analyst. Attributes that are familiar to and understood by respondents factor into, or influence, their choices, but those that are unfamiliar or not understood are less important in the decision process (Balbontin et al., 2017; Sandorf et al., 2017).

The no-choice option in a conjoint study allows respondents to choose something other than one of the brands included in the choice task (Brazell et al., 2006). Not only does the no-choice option serve to increase the realism of the respondent's decision, but it can also improve the resulting market share and volume predictions from the analysis (Carson et al., 1994) and eliminate statistical biases (Haaijer et al., 2001). By selecting the no-choice option, respondents are indicating a preference for an outside option, or the desire to opt out of the choice task altogether and save their money. This allows respondents to compare the utility of the alternatives in each choice task to some fixed level of utility they expect to achieve in the market (Louviere et al., 2010; Bahamonde-Birke et al., 2017).

Various psychological factors that claim to violate economic assumptions have been shown to be associated with the decision to select the no-choice option in behavioral studies (Dhar, 1997; Gunasti and Ross, 2008; Tversky and Shafir, 1992), including the types of alternatives that are more likely to lose choice share to no-choice options (Dhar and Simonson, 2003), and how the inclusion of a no-choice option changes the decision process (Parker and Schrift, 2011). Although the existing literature has focused on factors related to general choice deferral, we focus on the no-choice option in conjoint analysis. Consider, for example, a conjoint study of the determinants of elderly persons selecting an assisted living care facility. It is doubtful that many people are aware of different care options, or of the daily rate of assisted care, unless they are involved in the financing or care of another elderly person. The respondents probably have a much better grasp of the incremental value of increased services, such as better dining options. Since the selection of the no-choice option depends on the overall price levels of the inside good options, there is a chance that respondents make inferences about market prices from the choice tasks themselves when they are unfamiliar with the product category.

The effect of choice task uncertainty has been shown to produce contextual effects documented in the behavioral economics literature, such as attraction and compromise effects (Natenzon, 2019; Khaw et al., 2020; Enke and Graeber, 2023; He, 2024). Sometimes the uncertainty is due to the complexity of the choice task (Boxall et al., 2009) and sometimes it is due to imprecisely defined attribute-levels, such as stating that the performance of product along a specific dimension is "low," without providing a precise definition of what this means. Conjoint studies attempt to mitigate the effects of uncertainty by screening out respondents who are not familiar with the product category and by providing concise and concrete definitions of product features (Orme, 2010). Screening questions provide some degree of assurance that consumers are likely aware of product features and general price levels because of their past or intended purchase activities. However, there are instances when this is difficult to accomplish, such as when the entire product category is new to the market. In this setting, the "stable preference" assumption of conjoint analysis may not be true and respondents may learn about product features and prices from the conjoint exercise itself.

The purpose of this paper is to examine respondent learning about marketplace prices and product features in a conjoint exercise where the value of an offering is based on the concept of a random utility model (Louviere et al., 2010). If standard screening questions about product participation are sufficient for identifying qualified candidates who are aware of marketplace prices and product features, then providing additional pricing and attribute information will minimally affect part-worth estimates. However, if survey screening questions are an insufficient proxy for brand price and attribute knowledge, then providing price and attribute information may change consumer preferences for the outside good and possibly other product features. Invariance of consumer choices to outside good information is therefore fundamentally related to the assumption that the economic view of preference construction is valid for conjoint studies with properly screened respondents.

The organization of this paper is as follows. Section 2 describes a conjoint study in a new product category, provides model-free evidence of consumers updating preferences for the outside good and develops a model for the outside good. Section 3 describes a second study in an established product category in which respondents are exposed to alternative sets of information about market offerings and prices. Empirical results for the two studies are reported in Section 4, Section 5 provides a discussion of our results and concluding remarks are offered in Section 6.

2. MP3 Player Study

We begin our investigation using data from a MP3 player study conducted in 2008 when this product was new to the market and market prices were unlikely to be known by respondents. The seven attributes were chosen based on the technical innovations planned by the client for the upcoming year, and the prices were determined according to these new features for 2009. The brands reflect the current competition in the three countries. At the time the study was conducted, one of the key innovations in the market was the ability to play videos on MP3 players. Additionally, it's important to note that the iPhone was not yet available in these markets—it had only been announced for Germany and the UK in the following months. Table 1 reports the number of respondents from each country and the market penetration of the MP3 players in each at the time the study was fielded.

Table 1: MP3 Player Sample Size and Market Penetration

	France	Italy	UK
Number of Respondents	482	452	487
Market Penetration	3.4%	2.2%	5.3%

Fifteen sets of choice tasks, each with 15 tasks, were generated using a SAS algorithm that produces D-efficient, orthogonal and level-balanced experimental designs (Kuhfeld, 2010). The 15 sets of choice tasks were randomly assigned to the respondents and the order of the 15 choice tasks within these 15 sets was also randomized, so there would be no influence on the outside good due to the design. Before the experiment began, all attributes and variations were thoroughly presented and explained to the respondents through text and visual materials.

To provide respondents with context and aid their understanding of the market landscape, they were shown current MP3 player models available at the time. Awareness of these devices was moderate, with fewer than 5% of participants owning one. These existing devices typically offered storage capacities ranging from 4GB to 8GB, enabling users to store up to approximately 2,600 songs—a significant improvement compared to CD-based "Discman" players that could store only around 14 songs per disc. However, these early MP3 players lacked features such as satellite navigation (GPS), Bluetooth connectivity, touchscreens, or video playback. File transfer was limited to wired USB connections. In 2004, these devices were priced between £299 and £500.

Figure 1 displays MP3 player models from the four brands included in the study. Creative's device resembled a small box and lacked the aesthetic appeal typical of modern technology products. Samsung's model appeared similar to feature phones of the era, many of which included FM radios, although only a few high-end models supported MP3 playback and even then, with limited memory. Sony's player retained a design reminiscent of the Walkman and Discman product lines, signaling continuity with its legacy branding. In contrast, Apple introduced a distinctly innovative design with the iPod, departing from conventional aesthetics and setting a new standard in the category. While functionality across brands was broadly comparable, visual design served as a key point of differentiation.

Figure 1: MP3-Players in 2004

Before initiating the choice-based conjoint tasks, respondents were introduced to the upcoming generation of MP3 players (Figure 2). These models incorporated significant design and feature advancements, most of which were unfamiliar to the study participants. Although respondents were able to conceptually understand the new features based on the provided descriptions, they lacked experiential knowledge, making it difficult for them to fully evaluate the utility of these enhancements in practical, everyday use.

Figure 2: MP3-Players Planned for 2009

Interestingly, although the new models featured substantial technological improvements—including video capability, improved displays, and redesigned form factors—their prices remained within a similar range (£179–£599) as those of previous generations. The updated devices bore little resemblance to earlier models, adopting a more modern and functional tech aesthetic. Apple retained the iPod design, but enlarged the display to better support video content. In a noteworthy move, Apple also introduced a new device resembling an early version of the iPhone, albeit without mobile phone functionality.

The study was conducted in France, Italy and the United Kingdom, providing us the ability to replicate results across countries. There are four choice options including the outside good in each choice task. Table 2 reports the attributes and levels used in the study. There are four brands comprising the set of inside goods and a total of nine different attributes. Price is reported in units of GBP (Great British Pound), although tasks in each country were displayed in local currency. The first column of attribute-levels is listed parenthetically to indicate reference levels for each attribute. Part-worths for the remaining attribute-levels represent the marginal utility of the indicated level relative to the reference level for each attribute. The conjoint design was randomized across respondents so that different respondents were exposed to the choice tasks in a different order.

Table 2: MP3 Player Product Attributes and Levels

Attribute	Levels						
	1	2	3	4	5	6	7
Brand	Samsung	Apple	Sony	Creative	—	—	—
Shape	(t10)	iPod nano	iPod Touch	p2	—	—	—
LCD Display Size	(2.0")	3.0"	3.3"	—	—	—	—
Touch Screen	(No)	Yes	—	—	—	—	—
Mobile TV	(No)	Yes	—	—	—	—	—
Satellite Navigation	(No)	Yes	—	—	—	—	—
Wireless Connection	(None)	Wi-Fi	Bluetooth	Wi-Fi/Bluetooth	—	—	—
Storage Capacity	(4GB)	8GB	16GB	32GB	—	—	—
Price (GBP)	179	199	249	300	399	499	599

Note: The attribute levels in parentheses are used as the reference level of each corresponding attribute.

Model-free evidence of respondent updating their preference for the outside good is provided in Figure 3. Plotted is the proportion of choices for the outside good across the choice tasks for each country. A significant upward trend in the tendency of respondents to select the outside good is detected, indicating that the outside good does not exhibit constant utility across the choice tasks.

Due to the random task design assigned to each respondent, we can provide additional insight into the effect of respondent learning by dividing the dataset into two parts based on the task number. The first dataset consists of the first 8 tasks completed by each respondent (t = 1,…,8), while the second dataset consists of the last 7 tasks (t = 9,…,15). Figure 2 compares conjoint part-worth estimates from these datasets with a 45-degree line that indicates equality. Aside from the part-worths for the brands (Samsung, Apple, Sony, Creative), the results indicate equality of the part-worth estimates. We also find similar results from Italy and the UK. The part-worths for the brands are not on the 45-degree line and instead are plotted on a second parallel line in the figure. This implies that the brand part-worths are estimated to be similar relative to each other, but at a different overall level.

Our preliminary analysis indicates that respondents change their utility value for some (i.e., the brand coefficients) but not all (i.e., other product attributes and levels) of the conjoint model coefficients. In our proposed model, discussed below, we allow respondents to update only the outside good utility, while keeping the other attribute-level parameters constant. The result in Figure 4 shows that keeping other attribute-level parameters constant is a reasonable assumption while the values of the four brands drop relative to the increasing outside good drop. That finding supports the sufficiency of our proposed model.

Figure 3: Proportion choosing the outside good option in each task from the MP3 Player study.

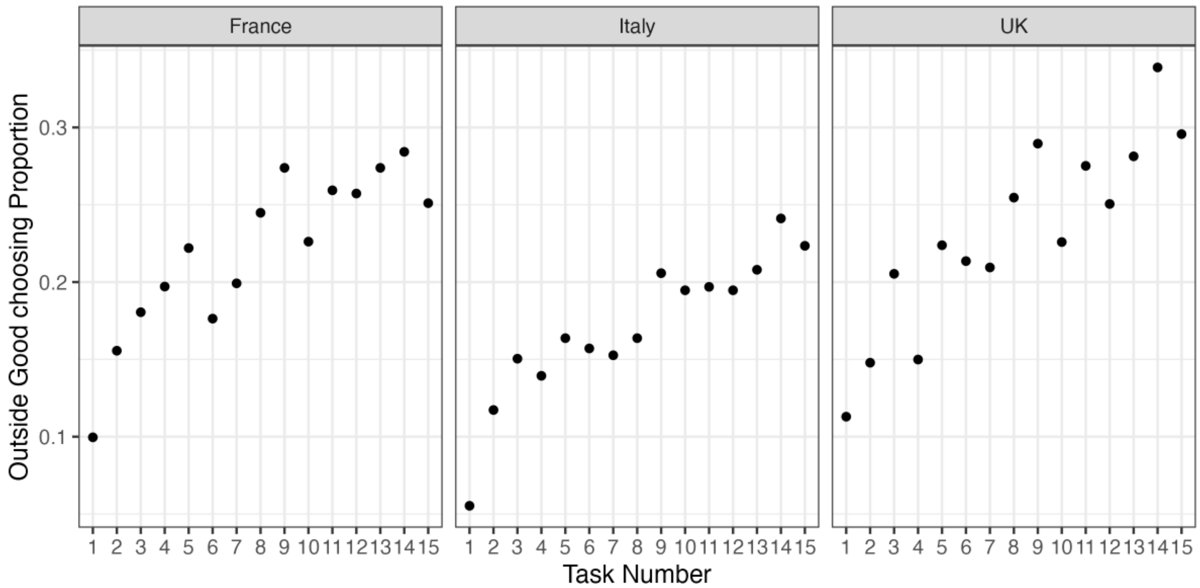
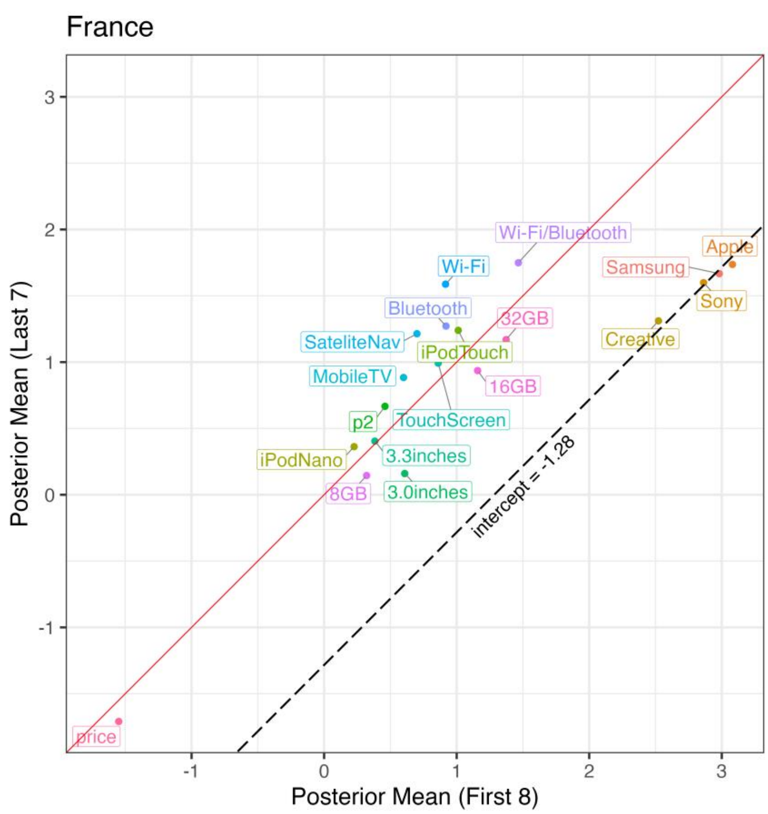


Figure 4: Comparison of mean of random-effect distribution of part-worths from the first 8 choice tasks (t = 1,…,8, x-axis) to last 7 tasks (t = 9,…,15, y-axis) from France.



3. TOOTH-WHITENING STUDY

Participants were recruited from a United States panel for the tooth-whitening study. This category was selected for analysis because tooth whitening products were first introduced into the market in the 1970's and these products are widely used today. Tooth-whitening products also exhibit large price variations and there are variety of product attributes. A list of product attributes used in the survey is presented in Table 3 along with their definitions.

Table 3: Tooth-Whitening Study Product Attributes, Definitions, and Levels

Attribute	Definition	Levels							
		1	2	3	4	5	6	7	8
Brand	The brand name of the teeth-whitening product	Crest	PlusWhite	Rembrandt	GoSmile	Auraglow	Luster	—	—
Form	The method of application of the product	(Strips)	Whitening Pen	Trays & Gel	Light Tech	—	—	—	—
Treatment Time	The total suggested time for a single whitening treatment	5	15	25	—	—	—	—	—
Number of Treatments	The total suggested number of whitening treatments to be completed	7	14	21	—	—	—	—	—
Time Until Results	The claimed amount of time until consumers should see visibly whiter teeth	3	7	14	—	—	—	—	—
Peroxide %	The percentage of the active ingredient (hydrogen peroxide) contained in the product	6	7	8	9	10	11	12	13
Price	The price of the product	14.99	24.99	34.99	44.99	54.99	—	—	—

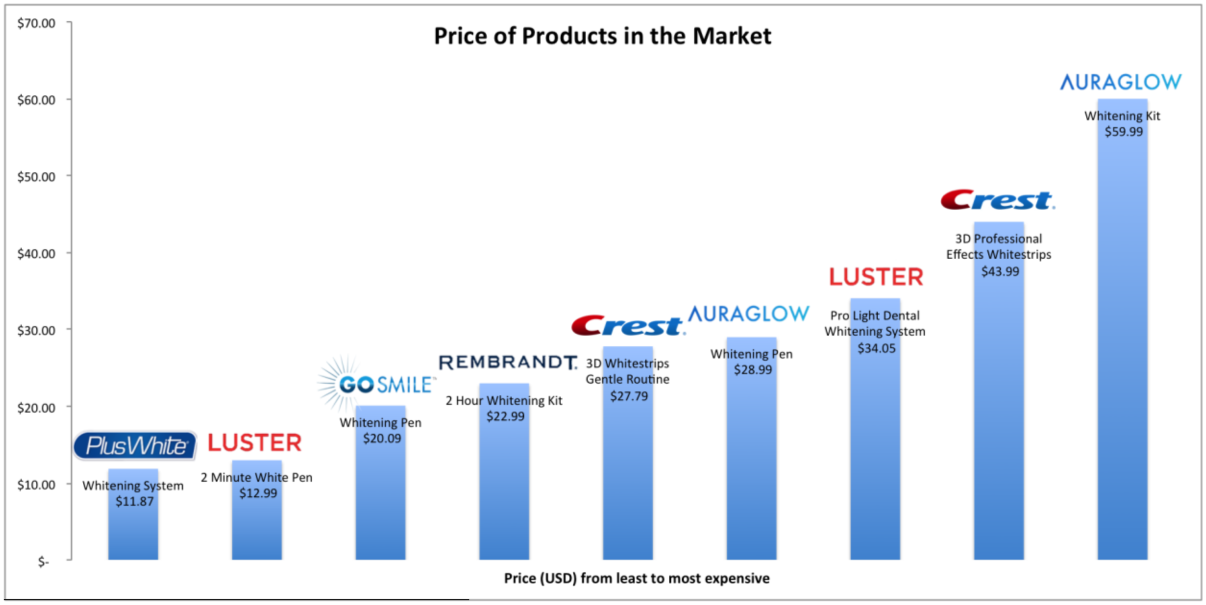
Note: The attribute levels in parentheses are used as the null level of each corresponding attribute.

Individuals responding to the Internet panel provider's invitation to participate were presented with a series of screening questions to determine whether the potential respondent was "in" the product category and had sufficient knowledge to provide informative answers to the survey questions. To be included in the survey, respondents must make their own hygiene purchases, be medically qualified to use the product, and be engaged in or actively considering buying some offering in the product category.

Qualified respondents who successfully passed the screening questions were then required to watch a short video describing the choice task and defining the product attributes and their levels. Respondents were then randomly assigned to one of three experimental conditions. The first condition did not provide any information about marketplace prices or attributes and serves as a control group for the other two conditions.

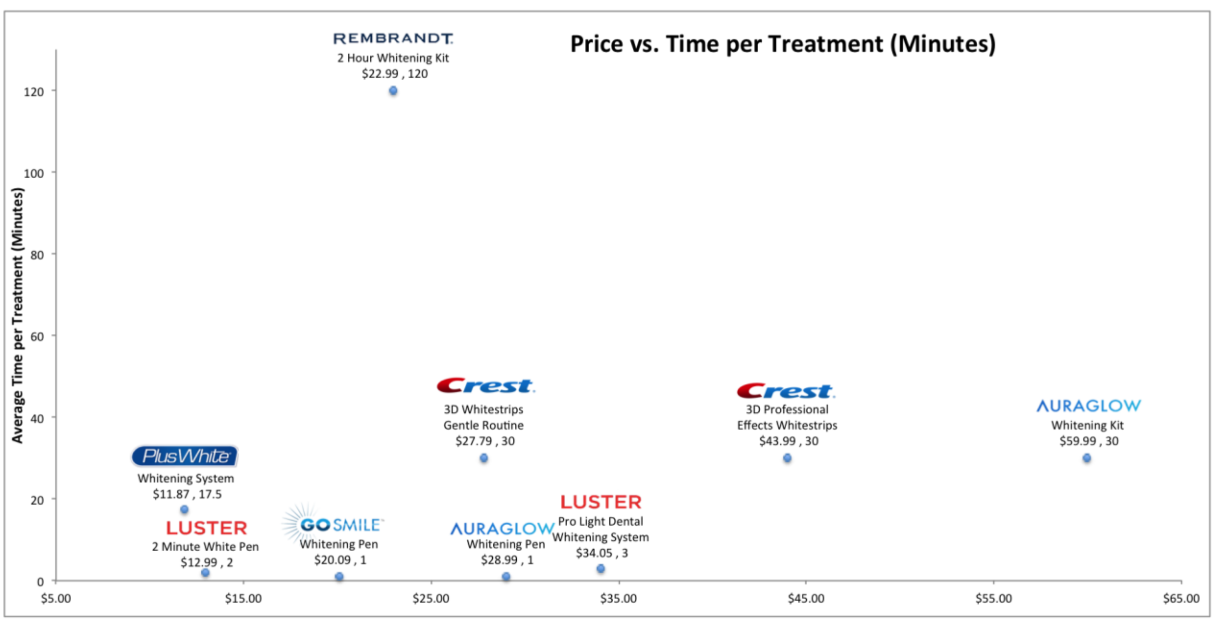
The second experimental condition provided respondents with the range of prices for the different brands under study as well as brands not included in the survey. Prices were obtained from the posted prices on Amazon.com. Figure 5 displays the information provided to respondents.

Figure 5: Price Distribution Graphic



The third experimental condition provided information about the relationship between brand, price, and the effectiveness attributes. Graphs comparing price to time per treatment, number of treatments, time to results, and percentage of peroxide were explained to respondents and show in succession. Figure 6 displays the price versus time per treatment graphic shown to respondents.

Figure 6: Price vs. Time per Treatment Graphic



Conjoint choice tasks were then presented to the qualified respondents, asking them to identify their most preferred offering from among those chosen. Each choice task included three different brand names and a no-choice option from which respondents indicated their preferred brand. Each respondent was asked to indicate their preferences across 12 choice tasks. An example choice task is shown in Figure 7.

The survey ended with a series of demographic variables to help assess the representativeness of the sample population. A total of 1,141 qualified respondents completed the survey and provided information about their preferences. The number of respondents was evenly split across the three experimental conditions.

Figure 7: Tooth Whitening Choice Task

	Option A	*Option B*	*Option C*	*I would not choose any of these.*
Brand	Crest	GO SMILE	Crest	
Form	Pen	Trays + Gel	Light Technology	
Time for One Treatment	5 minutes	15 minutes	15 minutes	
Number of Treatments	7	14	21	
Time to Results	3 days	7 days	3 days	
Percent Peroxide	12%	8%	10%	
Price	$14.99	$14.99	$34.99	-

4. Empirical Results

We compare three models for each of the datasets:

1. A standard HB choice model with heterogeneity.
2. An HB choice model with heterogeneity in which the value of the outside good is assumed to linearly increase in value across the choice tasks.
3. An HB choice model with heterogeneity in which the value of the outside good is probabilistic updated depending on the expected maximum utility of the inside goods.

The third model is referred to as our Proposed Model. It incorporates dynamic updating of, or learning about, the outside good with the amount of updating dependent on the value of the choice offerings in the previous choice task. Thus, the amount of updating is greater when respondents are exposed to a choice task where the options are more attractive.

Table 4 summarizes the model fit using the log marginal density (LMD) statistic. Values of LMD closer to zero indicate better model fit to the data.

Table 4: Model Fit

Study	Model	Dataset		
MP3		France	Italy	UK
	Standard model	-3932	-3891	-3965
	Linear model	-3767	-3756	-3818
	Proposed model	-3678	-3643	-3663
Tooth Whitening		Condition 1	Condition 2	Condition 3
	Standard model	-1656	-1766	-1973
	Linear model	-1584	-1702	-1920
	Proposed model	-1470	-1634	-1824

Tables 5 and 6 report the average part-worth estimates for the attribute levels, price, and the outside good updating parameter ρ of our proposed model. Estimates are reported in terms of the mean of the random effects distribution. There is general agreement in the estimates among the datasets in each study. This indicates that the MP3 study offers a valid setting for documenting the effect of the outside good on demand in a new product market, and the tooth-whitening study provides a basis for inferences about the degree of updating when respondents are confronted with differing amounts of information about the outside good in a more mature market.

Table 5: MP3 Posterior Mean (Standard Deviation) of Random Effects

Attribute	France	Italy	UK
Brand			
Samsung	4.62 (0.62)	5.26 (0.52)	6.15 (0.77)
Apple	4.70 (0.62)	5.30 (0.51)	6.49 (0.76)
Sony	4.54 (0.62)	5.33 (0.53)	6.39 (0.77)
Creative	4.30 (0.62)	5.12 (0.53)	5.80 (0.76)
Shape			
iPodNano	0.30 (0.13)	0.13 (0.12)	-0.01 (0.12)
iPodTouch	1.07 (0.15)	0.62 (0.12)	0.62 (0.12)
p2	0.47 (0.12)	0.04 (0.10)	0.11 (0.11)
LCD display size			
3.0inches	0.44 (0.09)	0.23 (0.09)	0.46 (0.09)
3.3inches	0.38 (0.09)	0.26 (0.09)	0.43 (0.09)
Touch Screen	0.86 (0.10)	0.83 (0.09)	0.77 (0.09)
Mobile TV	0.75 (0.09)	0.62 (0.09)	0.95 (0.10)
Satellite Navigation	0.90 (0.10)	0.87 (0.10)	0.72 (0.09)
Wireless Connection			
Wi-Fi	1.23 (0.13)	1.07 (0.12)	1.21 (0.12)
Bluetooth	1.01 (0.12)	0.98 (0.11)	1.10 (0.12)
Wi-Fi/Bluetooth	1.50 (0.13)	1.39 (0.12)	1.47 (0.13)
Storage Capacity			
8GB	0.31 (0.11)	0.33 (0.11)	0.44 (0.11)
16GB	0.99 (0.12)	0.75 (0.11)	1.20 (0.12)
32GB	1.24 (0.13)	1.13 (0.12)	1.59 (0.13)
Price	-1.56 (0.09)	-1.29 (0.07)	-1.71 (0.09)
ρ	-4.16 (0.23)	-4.30 (0.21)	-4.00 (0.21)

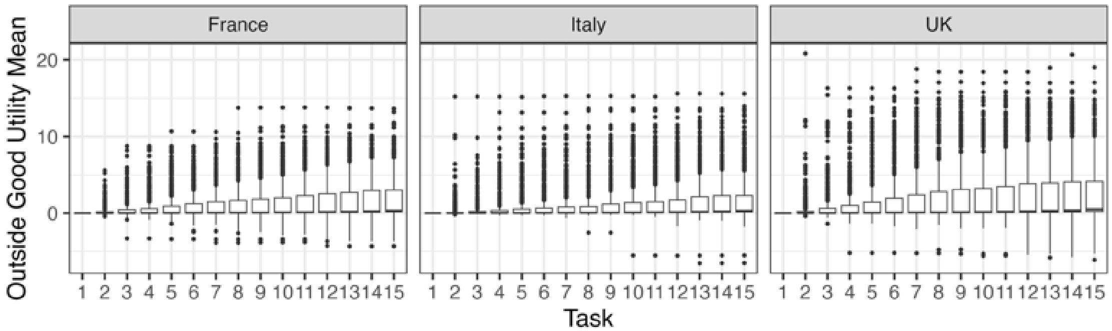
Table 6: Tooth-Whitening Posterior Mean (Standard Deviation) of Random Effects

Attribute	Condition 1	Condition 2	Condition 3
Brand			
Crest	34.13 (1.24)	32.34 (1.74)	28.30 (1.42)
PlusWhite	31.38 (1.12)	29.26 (1.61)	26.31 (1.37)
Rembrandt	33.39 (1.16)	30.71 (1.62)	28.02 (1.39)
GoSmile	32.60 (1.16)	31.33 (1.72)	27.83 (1.49)
Auraglow	32.06 (1.25)	30.78 (1.68)	27.80 (1.45)
Luster	31.51 (1.17)	29.62 (1.66)	26.82 (1.45)
Form			
Whitening Pen	0.12 (0.48)	-1.21 (0.52)	0.11 (0.46)
Trays & Gel	0.40 (0.53)	-0.40 (0.55)	0.51 (0.49)
Light Technology	-1.39 (0.51)	-1.39 (0.51)	-0.79 (0.45)
Treatment Time	-0.36 (0.04)	-0.33 (0.04)	-0.37 (0.03)
Number of Treatments	-0.17 (0.03)	-0.18 (0.03)	-0.15 (0.02)
Time Until Results	-0.19 (0.03)	-0.21 (0.03)	-0.25 (0.03)
Peroxide %	0.21 (0.08)	0.21 (0.07)	0.06 (0.06)
Price	-0.44 (0.03)	-0.35 (0.03)	-0.32 (0.02)
ρ	-5.06 (0.40)	-4.90 (0.46)	-5.01 (0.39)

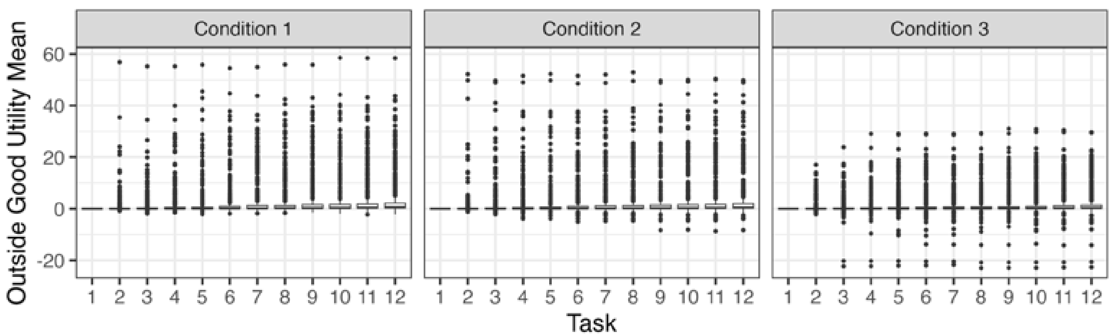
The primary difference in the estimates is in terms of the updating parameter ρ in Table 4 and the brand intercepts in Table 5. We find the updating parameter to be larger in the MP3 datasets than in the tooth-whitening datasets, indicating that respondents are updating their preference for the outside good to a greater extent as they progress through the survey. To illustrate the difference between the ρ estimates in the MP3 and tooth-whitening datasets, we calculate the update probability at least once by the end of the survey. This evaluates to 0.195 for France in the MP3 dataset and 0.067 for Condition 1 in the tooth-whitening dataset. In addition, the brand intercepts in the tooth-whitening datasets are larger in Condition 1 than in Condition 3, indicating that the additional information provided in Condition 3 increases the likelihood of selecting the outside good.

Figure 8 displays boxplots of the outside good utility for each respondent across choice tasks for both the MP3 and Tooth-whitening studies. The figure indicates that respondents in the MP3 player study are increasingly updating their outside good utility through the tasks, reflecting that they are learning about the market and the reference price when encountering new products. In contrast, respondents in the tooth-whitening study show only a small increase in outside good utility, indicating that updating occurs less often. The least amount of updating occurs in Condition 3 in which the respondents obtained the largest amount of market pricing and product information.

Figure 8: The individual outside good utility means for each respondent in (a) MP3 player study and (b) tooth-whitening study.



(a) MP3 player study



(b) tooth-whitening study

5. Discussion

The effect of the no-choice option has been discussed extensively in the behavioral decision theory (BDT) literature but has not been thoroughly investigated in the context of conjoint analysis where respondents are confronted with multiple choice tasks from which part-worths are estimated. Conjoint analysis differs from models used in BDT research where attraction effects (Dhar, 1997) and choice uncertainty (Dhar and Simonson, 2003) have been shown to affect preferences for the no-choice option.

In a conjoint analysis, attributes and their levels are defined in a concise and concrete manner (Orme, 2010), seeking to avoid the presence of task uncertainty. The instructions for expressing preferences in conjoint tasks ask respondents to imagine that unmentioned attributes are the same for each choice option and therefore do not factor into the relative preferences for the choice options [see Gunasti and Ross (2008)]. In addition, inferences about part-worths are made based on a dozen or more choice tasks, so that any contextual effect that may influence choice in one task is different than those in another choice task and are treated as part of the error term of the model. Therefore, a respondent wishing to avoid being too extreme in their expressed preferences, for example, will not have this tendency expressed in any specific part-worth (Simonson and Tversky, 1992). Similarly, attraction and dominance effects (Busemeyer and

Townsend, 1993; Huber et al., 2014) will not be reflected in a particular part-worth because of multiple choice tasks. Conjoint studies also screen out respondents who are not familiar with the offerings in a product category, whereas BDT studies often do not employ screens. Respondents who pass the screening questions are assumed to be already familiar with the product features and the price of products charged in the marketplace, and they can recall this information without being affected by the context of the choice task.

We find in our analysis that preference for the no-choice option is dynamically updated in the MP3 dataset where all offerings are new to the market and respondents lack a full understanding of the product and its benefits. Dynamic updating is still present in the teeth-whitening dataset but is significantly reduced and only present in the first two treatment conditions. This is consistent with a mature category wherein standard screening yields respondents with greater knowledge of prices and product offerings. In the third treatment condition these market details are clearly communicated before the conjoint study, so learning no longer occurs through the tasks and dynamic updating is eliminated.

We find from our analysis of the MP3 data, where respondents are less familiar with market prices and offerings, that preference for the no-choice option is dependent on the prices and features of the inside good options. In contrast to updating based on learning the trade-off between quality and price, as described in Simonson and Tversky (1992), we argue that respondents understand the relative differences between attribute levels (e.g., different display sizes), but not the no-choice option. In conjoint analysis, the brand part-worths are measured relative to outside good, which is set equal to zero in our analysis. This finding, that the part-worths for attribute-level are unaffected while the brand part-worths are affected by uncertainty about the market, is reflected in the model-free evidence shown in Figure 4.

More importantly, we find that the price coefficient estimate is unaffected, indicating that respondents are using the price locally to compare alternative inside goods, but not globally. That is, they allow price to impact their choice within a subset of available options (secondary demand) but do not consider the broader implications of price on overall market affordability (primary demand).

An implication of our findings relates to the design of choice-based conjoint tasks. Our results suggest that researchers and practitioners should consider the potential impact of learning when including the no-choice option, even when the inside good options are hypothetical. We argue that even for familiar products like tooth-whitening products, using appropriate video warm-ups and providing additional market information can help mitigate the updating phenomenon across tasks.

6. Conclusion

This paper revisits the no-choice option in conjoint analysis and finds that preference for it is affected by the respondent's familiarity with the product category. When the general level of prices is not known, as in the MP3 dataset, respondents update their preference for the outside good as they learn about the distribution of market prices. In the teeth-whitening dataset, dynamic updating is significantly reduced, in part due to consumer familiarity with the category. All dynamic updating is effectively eliminated when information about products and prices in the market is shared prior to the study. We present evidence that non-brand part-worth estimates are unaffected by dynamic learning, while the brand part-worths are affected. This implies that it is useful to remind respondents in a conjoint setting of the prices of products available in the market.

The important findings of our analysis are as follows. Our analysis:

1. Argues that choice-based conjoint should include the no-choice option. Otherwise, the predicted sales effect of a price or feature change will be confounded by changes in preference for the unmeasured alternative.
2. Provides evidence that the value of the outside good is stable in a study where respondents are familiar with the product category and understand the personal implications of using the brands and attributes.
3. Indicates that, for the well-understood Tooth-whitening category, presenting more information in terms of competing information has little impact on stabilizing preference for the outside good.
4. Demonstrates, for the MP3 analysis, a clear example for a new product category where the value of the outside good substantially increases and that the brands lose value across tasks.

Our results reinforce the idea that a properly designed and executed conjoint study, where respondents are screened for inclusion based on product familiarity, product features are described in concrete and unambiguous terms, respondents are instructed to assume missing attributes are the same for all choice alternatives, and a training video is presented to respondents so that they are familiar with the choice task, is largely immune to potential biases identified in laboratory experiments that lack these qualities. The no-choice option is needed in conjoint analysis because of the limited number of choice alternatives that can be presented to respondents in a choice task and carries with it the implicit assumption that respondents are aware of the prices charged for choice alternatives in the marketplace. Informing respondents of these prices prior to the conjoint choice task is encouraged but may not be necessary if prices are generally known. We encourage additional research in this area.

Cheng-Yu Hung

Peter Kurz

Roger A. Bailey

Joel Huber

Greg M. Allenby

References

Bahamonde-Birke, Francisco J., Isidora Navarro, Juan de Dios Ortúzar. 2017. If you choose not to decide, you still have made a choice. *Journal of Choice Modelling* **22** 13–23.

Balbontin, Camila, David A. Hensher, Andrew T. Collins. 2017. Do familiarity and awareness influence voting intention: The case of road pricing reform? *Journal of Choice Modelling* **25** 11–27.

Boxall, Peter, Wiktor L Adamowicz, Amanda Moon. 2009. Complexity in choice experiments: choice of the status quo alternative and implications for welfare measurement. *Australian Journal of Agricultural and Resource Economics* **53**(4) 503–519.

Brazell, Jeff D, Christopher G Diener, Ekaterina Karniouchina, William L Moore, Válerie Séverin, Pierre-Francois Uldry. 2006. The no-choice option and dual response choice designs. *Marketing Letters* **17**(4) 255–268.

Busemeyer, Jerome R, James T Townsend. 1993. Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review* **100**(3) 432.

Carson, Richard T., Jordan J. Louviere, Donald A. Anderson, Phipps Arabie, David S. Bunch, David A. Hensher, Richard M. Johnson, Warren F. Kuhfeld, Dan Steinberg, Joffre Swait, Harry Timmermans, James B. Wiley. 1994. Experimental analysis of choice. *Marketing Letters* **5**(4) 351–367.

Dhar, Ravi. 1997. Consumer Preference for a No-Choice Option. *Journal of Consumer Research* **24**(2) 215–231.

Dhar, Ravi, Itamar Simonson. 2003. The effect of forced choice on choice. *Journal of Marketing Research* **40**(2) 146–160.

Enke, Benjamin, Thomas Graeber. 2023. Cognitive uncertainty. *The Quarterly Journal of Economics* **138**(4) 2021–2067.

Gunasti, Kunter, William T. Ross. 2008. How Inferences about Missing Attributes Decrease the Tendency to Defer Choice and Increase Purchase Probability. *Journal of Consumer Research* **35**(5) 823–837.

Haaijer, Rinus, Wagner Kamakura, Michel Wedel. 2001. The "no-choice" alternative to conjoint choice experiments. *International Journal of Market Research* **43** 93–106.

He, Junnan. 2024. Bayesian contextual choices under imperfect perception of attributes. *Management Science* **70**(3) 1465–1482.

Huber, Joel, John W Payne, Christopher P Puto. 2014. Let's be honest about the attraction effect. *Journal of Marketing Research* **51**(4) 520–525.

Khaw, Mel Win, Ziang Li, Michael Woodford. 2020. Cognitive Imprecision and Small-Stakes Risk Aversion. *The Review of Economic Studies* **88**(4) 1979–2013.

Kuhfeld, Warren F. 2010. *Marketing Research Methods in SAS.* SAS Institute Incorporated. https://support.sas.com/resources/papers/tnote/tnote_marketresearch.html.

Louviere, Jordan J, Terry N Flynn, Richard T Carson. 2010. Discrete choice experiments are not conjoint analysis. *Journal of Choice Modelling* **3**(3) 57–72.

Manski, Charles F, Daniel McFadden, et al. 1981. *Structural analysis of discrete data with econometric applications*. Mit Press Cambridge, MA.

Natenzon, Paulo. 2019. Random choice and learning. *Journal of Political Economy* **127**(1) 419–457.

Orme, Bryan K. 2010. *Getting started with conjoint analysis: strategies for product design and pricing research.* Madison: Research Publishers LLC.

Parker, Jeffrey R., Rom Y. Schrift. 2011. Rejectable choice sets: How seemingly irrelevant no-choice options affect consumer decision processes. *Journal of Marketing Research* **48**(5) 840–854.

Sandorf, Erlend Dancke, Danny Campbell, Nick Hanley. 2017. Disentangling the influence of knowledge on attribute non-attendance. *Journal of Choice Modelling* **24** 36–50.

Simonson, Itamar, Amos Tversky. 1992. Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research* **29**(3) 281–295.

Tversky, Amos, Eldar Shafir. 1992. Choice under conflict: The dynamics of deferred decision. *Psychological Science* **3**(6) 358–361.

OPTIMIZING PRODUCT PORTFOLIOS WITH DISCRETE CHOICE MODELS: A PRACTICAL APPROACH FOR MARKET RESEARCHERS

MAXIMILIAN RAUSCH
PETER KURZ
STEFAN BINNER
BMS MARKETING RESEARCH + STRATEGY

INTRODUCTION

Product portfolio optimization is a complex yet essential task in fast-moving consumer goods (FMCG) markets. With limited shelf space, intense competition, and heterogeneous customer preferences, managers must frequently ask themselves:

Figure 1

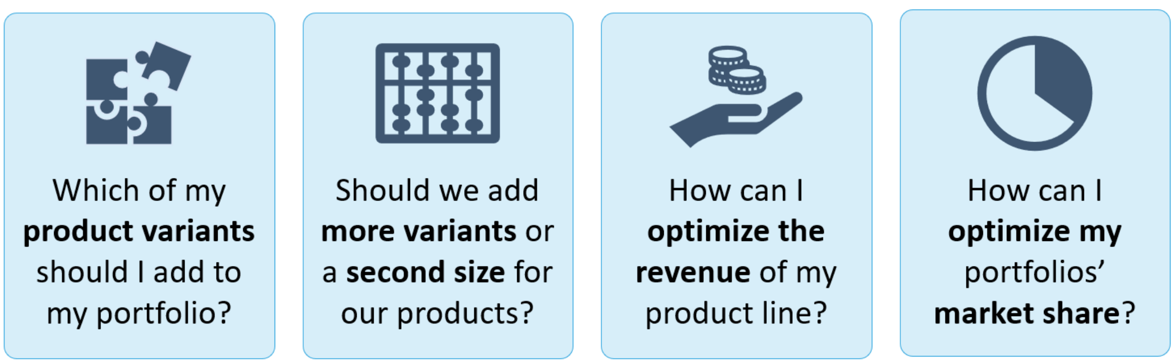


Figure 1 outlines the key dimensions relevant to portfolio optimization. In most real-world cases, factors such as assortment size, number of products, sales volumes, and pricing interact in complex ways. Consequently, portfolio decisions often involve balancing multiple objectives, necessitating either separate or combined target functions. This complexity can result in divergent optima or introduce constraints that must be explicitly considered when defining the optimal product mix. These challenges reflect a central insight: strategic priorities—such as revenue maximization versus market share growth—typically demand distinct portfolio configurations.

FMCG markets are typically characterized by high product differentiation, complex brand architectures (main brands, sub-brands), and numerous stock-keeping units (SKUs). A retailer like Walmart may carry over 400 detergent SKUs, reflecting vast consumer heterogeneity. In such environments, manufacturers must design portfolios that not only appeal to diverse consumer segments but also meet the practical constraints imposed by retailers.

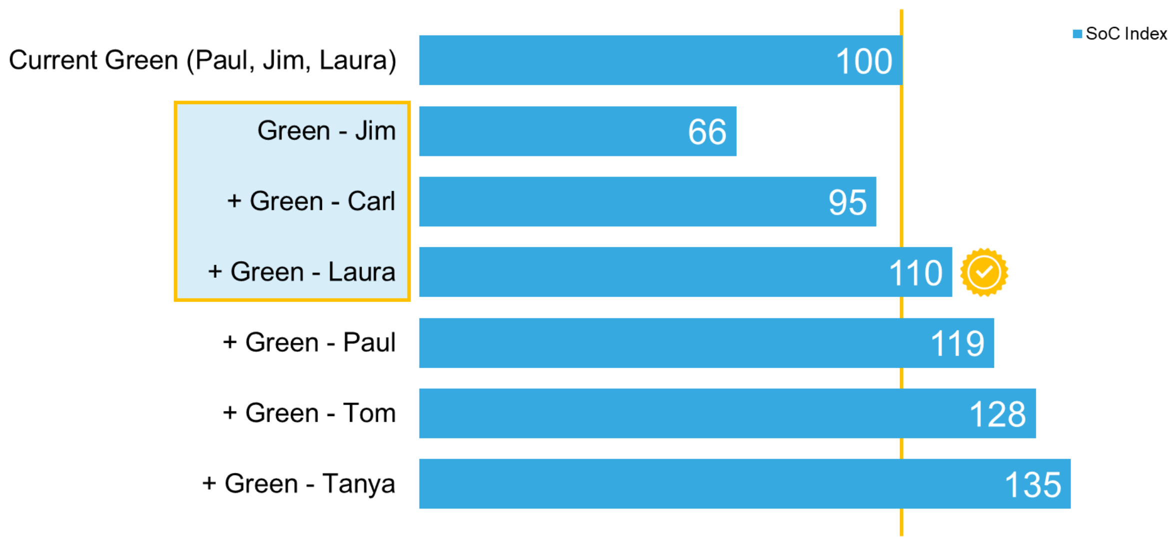
This paper proposes a portfolio optimization framework that leverages part-worth utilities derived from conjoint analysis. We demonstrate how these utilities can be integrated into various optimization techniques—including exhaustive search, genetic algorithms (GA), and simulated annealing (SA)—to identify product portfolios that perform strongly against predefined strategic objectives.

Let's Start Simple

For complex choice modeling studies, it is essential to ensure good data quality. Especially when the goal of the study is to optimize more complex portfolios, the sample size needs to be large enough, bad respondents need to be removed, and it is key that the estimation converges.

To find the best set of X out of Y products several approaches can be followed. One simple approach is to start with a clean slate and iteratively add the next best SKU. Start with the best single SKU and then stepwise repeat this process until the desired portfolio size is reached. This algorithm returns a hierarchy of the products; it however ignores that SKU combinations of different ranks might outperform earlier combinations. Previous combinations are fixed in this approach.

Figure 2



In our example in Figure 2 we look at a fictitious brand "Green" with 6 potential SKUs. We can see that the stepwise approach reveals that the three SKUs "Jim," "Carl" and "Laura" reach a higher share of choice (SoC) compared to the current three SKUs. "Carl" does replace "Paul" in this optimized scenario of a three SKU portfolio.

This easy algorithm delivers quick results even for complex studies with larger sets of SKUs. It however fixes the previous SKUs in each step which ignores potential better combinations in larger portfolios. To tackle this more systematically and in more depth a different approach will help.

We can define a fixed competitive context and then simulate all possible combinations of our client's SKUs. This long list of potential assortments can then be filtered through by applying the relevant business considerations.

- Include any must have SKUs
  Maybe there are some SKUs which define the client's core brand values and need to be included in any assortment.

- Consider product margins and production costs
  Some SKUs might cost more in production and hence the margins will be lower.

We will end up with a shortlist of potential assortments which you can then evaluate in detail and find the optimal business case for the task at hand.

Running all combinations is feasible for smaller problems.

$$\binom{6}{1}+\binom{6}{2}+\binom{6}{3}+\binom{6}{4}+\binom{6}{5}+\binom{6}{6}=2^6-1=63$$

The simple example of 6 potential SKUs ends up with 63 possible assortment combinations. For 10 SKUs it will already be 1023 and with 20 the limit of lines in MS Excel is reached (2 to the power of 20). The complexity needs to be reduced.

Reducing the complexity can be achieved by some practical considerations.

- Groups of SKUs that can be combined and optimized separately
- Must-have SKUs that have to be in an assortment
- Minimum and/or maximum number of SKUs in the portfolio

Applying these business related rules does usually reduce the complexity of the search space for combinations.

To get to the optimal solution for the resulting optimization problem mathematical optimization methods (e.g., simulated annealing, genetic algorithms or exhaustive search) can be used. In the following we describe and evaluate three mathematical optimization algorithms that can be applied to conjoint study results.

Portfolio Optimization Techniques

Portfolio optimization based on conjoint data begins with the estimation of part-worth utilities for attribute levels derived from a choice-based conjoint (CBC) experiment. These utilities serve as the foundation for simulating market scenarios and assessing the performance of different product configurations.

A common objective is to maximize preference or market share. The most straightforward assumption follows a deterministic choice rule, where the product combination with the highest utility is always chosen. However, as highlighted by Green and Krieger (1988, 1992), this assumption may be overly simplistic due to estimation errors, variability in consumer behavior, and the presence of low-involvement or highly similar alternatives. In such contexts, consumers may not consistently select the utility-maximizing option.

To address these limitations, probabilistic choice rules are widely applied in the conjoint literature. These models assign choice probabilities to product alternatives while preserving the ordinal structure of utility estimates. A particularly prevalent model is the logit rule (McFadden 1976; Punj and Staelin 1978), which assumes that the probability of a product being chosen is proportional to the exponentiated utility. This probabilistic framework is adopted in the present study to reflect more realistic consumer behavior patterns.

Effective portfolio optimization requires the formulation of a target function that evaluates the performance of candidate portfolios. This involves selecting the actual products to include on the shelf, determining the optimal number of products per product line, and assembling product profiles based on the estimated part-worth utilities. The form of the target function depends on the strategic objective—e.g., maximizing revenue, preference share, or profit. Formula 1 illustrates an example of a probabilistic choice rule used for calculating preference share as the optimization criterion.

$$M = \frac{\sum_{j=1}^{E} \exp(\sum_{k=1}^{K} \sum_{l=1}^{L_k} \beta_{ikl} x_{jkl})}{\sum_{j'=1}^{J+E} \exp(\sum_{k=1}^{K} \sum_{l=1}^{L_k} \beta_{ikl} x_{j'kl})} \qquad \forall i, j, k, l \longrightarrow \max! \tag{1}$$

K = 1, …, K:	Number of Attributes
l = 1,…, Lk:	Number of Levels per Attributes
j = 1,…, E:	Number of Products per Product line
j = 1,…, J:	Actual Products in Shelf

In this approach, the existing products on the shelf serve as the starting point. The optimization is based on their attribute levels and the number of products included. To ensure meaningful outcomes, appropriate constraints must be applied. Without such restrictions, the model might suggest unrealistic portfolios—for instance, one where all products are offered at the lowest price, which would trivially maximize preference share.

$$\sum_{l=1}^{L_k} x_{jkl} = 1 \qquad \forall j, k, \tag{2}$$

$$\sum_{l=1}^{L_k} x_{jkl} = \sum_{l=1}^{L_{k+1}} x_{j(k+1)l} \qquad \forall j, k = 1, ..., K-1, \tag{3}$$

$$x_{jkl} \in \{0, 1\} \qquad \forall k, l, j. \tag{4}$$

Key constraints and parameters include:

- The number of distinct product lines (or portfolios) to be optimized (2),
- The total number of feasible product variants available in the market, reflecting the size of the theoretical choice set (3),
- The number of products that should be included within each product line (4).

These constraints help define a realistic search space for optimization and ensure the resulting portfolio configurations are both practical and implementable.

While this section focuses on optimizing preference share, the framework can be extended to alternative objectives. For revenue maximization, price and expected sales volume must be incorporated into the target function. If the goal is to maximize profit, unit costs must also be considered. Each optimization objective thus requires a tailored formulation of the target function, highlighting the strategic relevance of aligning the optimization criteria with the overarching business goals.

Depending on the complexity, we apply one of three primary optimization techniques:

- **Exhaustive Search**
  Compute all possible combinations and sort them in descending order by the optimization criteria (optima is guaranteed).
- **Simulated Annealing**
  Motivation comes from annealing in metallurgy, a technique involving heating and controlled cooling of material to increase the size of its crystals and reduce their defects (R-package optim with “SANN”).
- **Genetic Algorithms**
  Simulation of Selection (survival of the fittest), recombination (crossover) and mutation (variation) like in the evolution (R-package GA; also implemented in Lighthouse Studio).

Each method has distinct advantages and trade-offs with respect to speed, optimality, and scalability.

EXHAUSTIVE SEARCH

Exhaustive search entails the generation and evaluation of all possible combinations of product variants, ranking them according to a predefined objective function (e.g., expected revenue, preference share). This brute-force approach guarantees identification of the global optimum and thus serves as a valuable benchmark for heuristic optimization methods.

However, the method is computationally intensive and scales poorly with increasing problem complexity. As the number of attributes and levels grows, the number of potential portfolios expands combinatorially, making full enumeration infeasible for larger problem spaces. To strike a balance between optimality and practicality, we limit evaluation to the top 1,000 portfolio configurations, which are then further assessed based on business-specific criteria such as brand consistency, production feasibility, or profit margins.

Advantages:

- Guarantees the global optimum
- Provides a benchmark for evaluating heuristic methods
- Useful for multi-criteria decision-making and scenario analysis

Limitations:

- Computationally infeasible for large-scale problems
- Subject to combinatorial explosion with increasing complexity

Simulated Annealing (SA)

Simulated annealing is a probabilistic metaheuristic inspired by the annealing process in metallurgy. It approximates the global optimum by iteratively exploring the solution space while allowing occasional acceptance of suboptimal solutions. This mechanism helps avoid entrapment in local optima—a key advantage in high-dimensional optimization problems such as portfolio selection.

Core components of the algorithm include:

- **Initialization:** The process starts with a randomly generated portfolio or a base case reflecting the current market configuration.
- **Perturbation:** A candidate portfolio is generated by slightly modifying the current portfolio (e.g., by changing a product's attribute level or replacing a SKU).
- **Evaluation:** The candidate solution is evaluated using the predefined target function (e.g., preference share, revenue, or profit).
- **Acceptance Rule:** If the candidate is superior, it replaces the current solution. If it is inferior, it may still be accepted with a probability based on the Boltzmann distribution.
- **Cooling Schedule:** Over time, the probability of accepting worse solutions decreases, guiding the algorithm from exploration toward exploitation.

This process results in a **trace**—a sequence of intermediate solutions—which often includes several high-performing portfolios close to the global optimum.

Figure 3

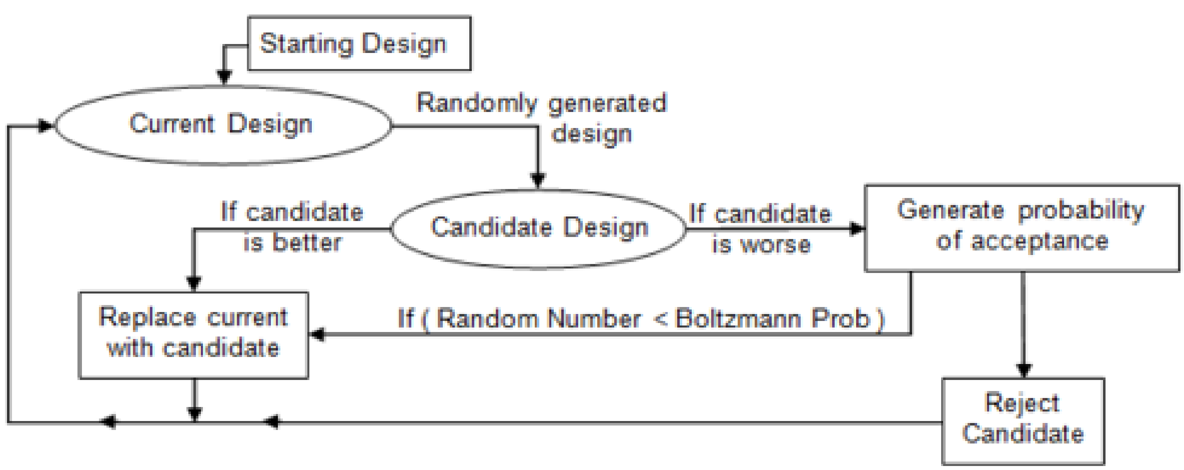


Figure 3 illustrates the cooling process, including the probabilistic acceptance of inferior solutions. The core idea is to maintain diversification in early stages of the search and gradually narrow down toward optimal regions of the solution space.

Advantages of Simulated Annealing:

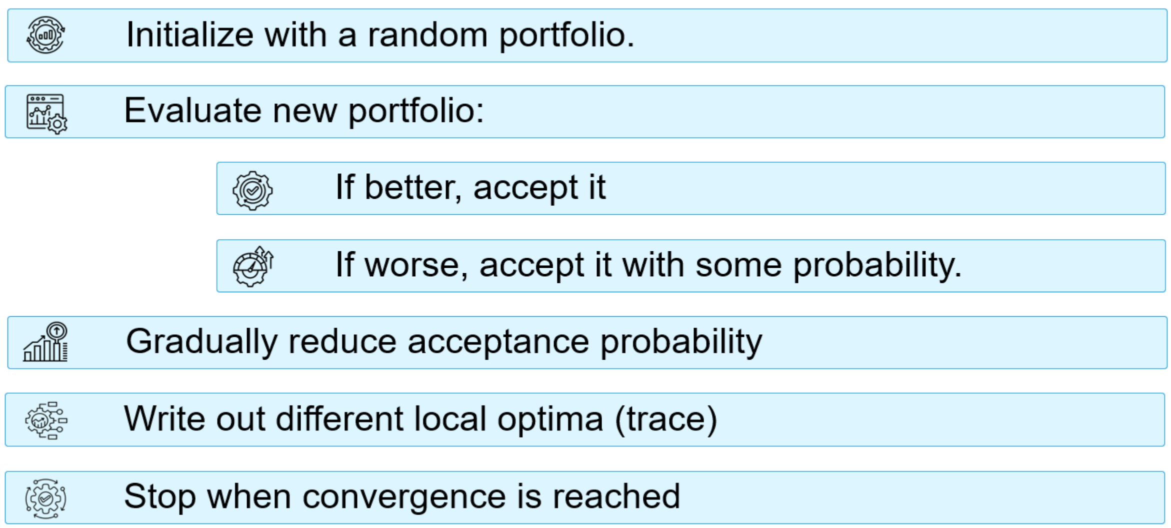
- Suitable for complex, high-dimensional problem spaces

- Balances global exploration and local exploitation
- Produces a diverse set of near-optimal portfolio candidates

From a practical standpoint, simulated annealing is particularly useful in portfolio optimization domains characterized by **NP-hard complexity**, where deterministic or exhaustive methods become computationally infeasible. The strength of SA lies in its flexible balance between accepting and rejecting suboptimal solutions—a mechanism that mitigates the limitations of both pure random search and strict gradient-based optimization.

Figure 4 presents a flow diagram of the simulated annealing process. The algorithm continuously perturbs and evaluates candidate portfolios, adjusting the search behavior based on the acceptance criteria until convergence is achieved.

Figure 4



Researchers may initiate the simulated annealing process either with a randomly generated portfolio or by using an existing market scenario as the baseline. In each iteration, the algorithm introduces small changes to the current solution—for example, by exchanging SKUs or modifying attribute levels—to generate a new candidate portfolio.

The candidate is then evaluated using the specified target function (e.g., preference share, revenue, or profit). If the new solution outperforms the current one, it is accepted as the new baseline. If it performs worse, it may still be accepted with a probability determined by the current "temperature" in the algorithm's cooling schedule, following the Boltzmann distribution.

Depending on whether the candidate is accepted or rejected, the next iteration proceeds either from the improved design or from a further perturbed version of the previous solution. This iterative process continues until convergence is observed—that is, until no significant improvement in the target function is detected over a defined number of iterations.

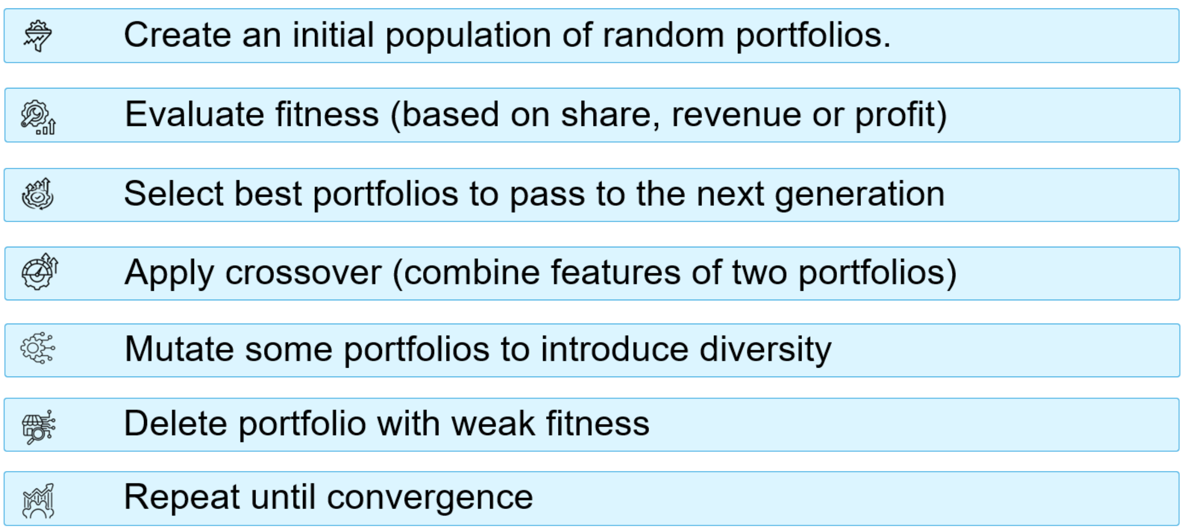
Genetic Algorithms (GA)

Genetic algorithms simulate the process of biological evolution by representing candidate solutions—such as product portfolios—as chromosomes composed of individual genes, which correspond to attribute levels or product features. The algorithm evolves these portfolios across multiple generations using mechanisms inspired by natural selection: **selection**, **crossover**, and **mutation**.

The process typically begins with an **initial population** of randomly generated portfolios. In each generation, a new population is created by selecting high-performing solutions from the current generation and applying genetic operations (Figure 5).

Process Overview

Figure 5



This evolutionary cycle is repeated until a stopping criterion is met—such as a maximum number of generations, convergence in the target function (certain number of generations without an improvement in solution quality), or a threshold level of solution quality.

Genetic algorithms can vary significantly in their implementation. Differences exist in how the population is initialized and maintained, the specific selection strategies used (e.g., tournament selection, roulette wheel), and how crossover and mutation are applied. These choices influence both the efficiency and the diversity of the search process.

Due to their flexibility and robustness, GAs are particularly suitable for complex portfolio optimization problems with large, nonlinear, and discrete solution spaces where traditional methods may fail to find satisfactory solutions.

Advantages:

- Fast convergence toward optimal solutions
- Avoids local optima via mutation

- Scalable to large problem spaces

However, genetic algorithms often require substantial computational effort and careful tuning of their parameters to achieve satisfactory results (for a detailed discussion of parameter effects, see Bayer and Voekler, 2023).

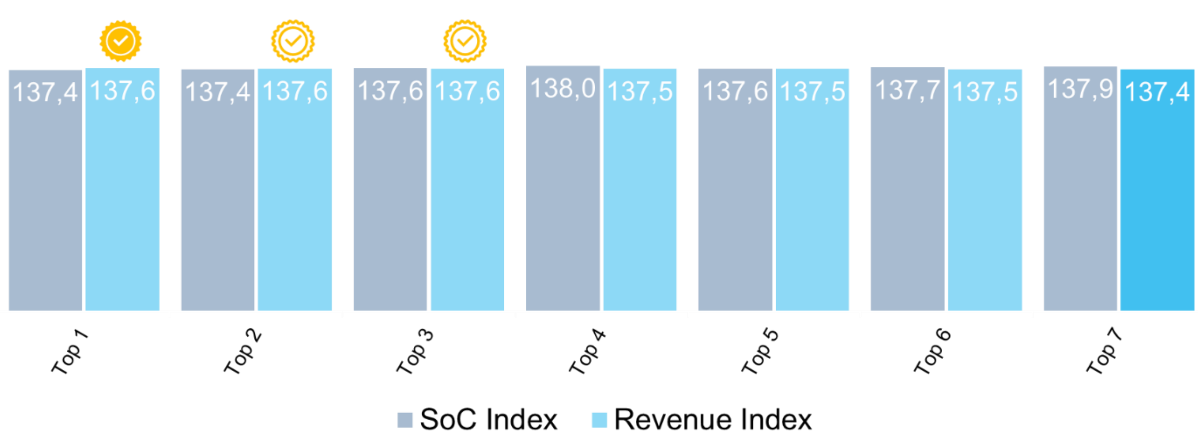
Let's Get Collaborative!

While the mathematical algorithms discussed can identify solutions that are optimal in terms of the defined objective functions, these **mathematical optima do not always align with business reality**. A solution that scores highest on predicted preference share, revenue, or profit may still be impractical due to factors such as brand architecture, channel strategy, production feasibility, or competitive dynamics.

Moreover, in more complex optimization problems, the top-ranked portfolio configurations often yield **very similar KPI values**. This means that there is typically not a single "best" solution, but rather a set of **high-performing alternatives** that perform nearly identically in quantitative terms.

Figure 6 illustrates this point by showing the top 7 portfolio configurations identified in the example presented below. While they differ in composition, their performance—measured by the chosen objective function—is nearly indistinguishable. This underlines the importance of applying **managerial judgment and qualitative evaluation** in selecting the final solution from among the top candidates.

Figure 6



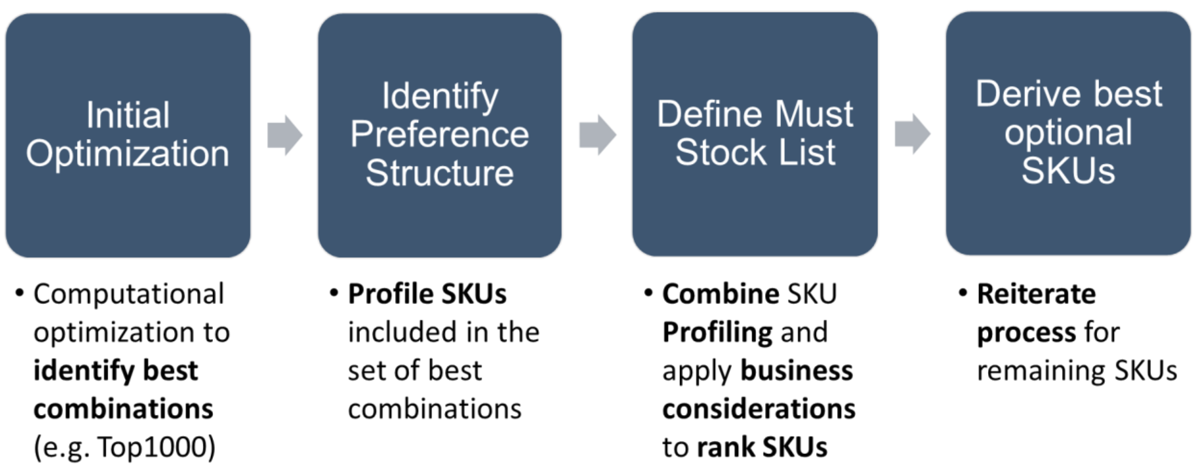
Given the typically small performance differences among top-ranked solutions, our approach does not focus on identifying a single "best" portfolio. Instead, we analyze a **set of high-performing configurations**, usually the top 1,000 combinations as determined by the optimization algorithm.

Within the **collaborative optimization framework** we propose, this top set serves as the starting point for deeper analysis. Specifically, we examine the **preference structure** by profiling the SKUs that appear most frequently across these high-performing portfolios. The resulting frequency distribution allows us to identify the **most preferred and strategically relevant SKUs**—those that consistently contribute to strong portfolio performance.

By combining these quantitative insights with **business-specific considerations**—such as brand strategy, product fit, and operational feasibility—we can derive a **prioritized ranking** of SKUs for inclusion in the final portfolio.

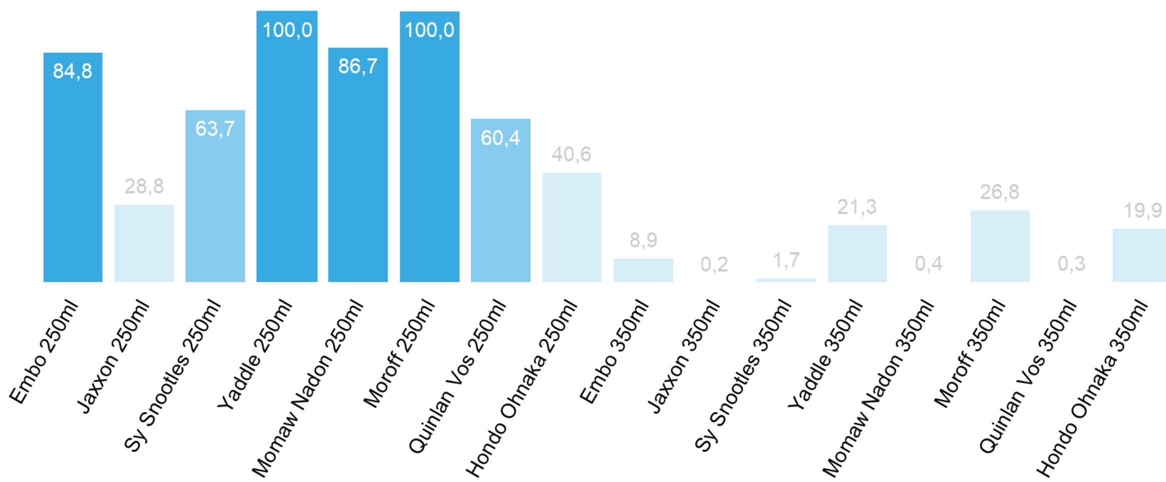
As an optional final step, the process can be **iterated** to assess the remaining SKUs not included in the top-performing portfolios. The necessity of this step depends on the cumulative performance achieved by the initially selected SKUs. If a sufficiently high share or revenue is already captured, further refinement may be unnecessary; otherwise, the iterative process can help identify additional, second-tier SKUs to complete the portfolio.

Figure 7



To illustrate this approach we will look at an example of an anonymized study. This study has different SKU variants at different sizes. In Figure 8 we show the 250ml and 350ml options.

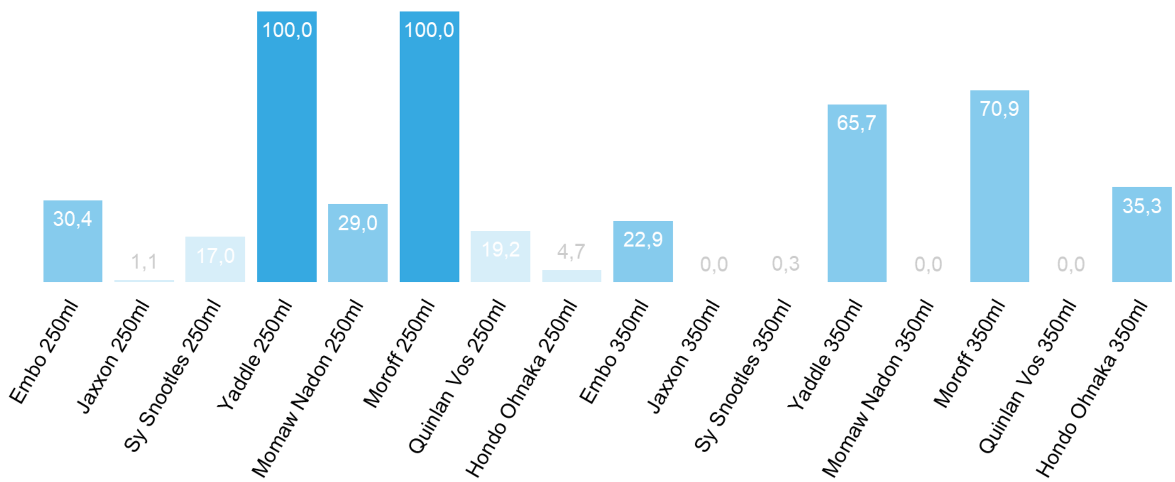
Figure 8



The chart displays the **inclusion frequencies** of individual SKUs across the top 1,000 portfolio combinations. It illustrates how often each SKU appears within these high-performing configurations. Notably, the SKUs *"Yaddle"* and *"Moroff"* are present in all 1,000 combinations, indicating their strong and consistent contribution to portfolio performance. This suggests that these SKUs are particularly relevant and likely represent core preferences within the target market.

Furthermore, the distribution of frequencies reveals a general **preference for 250ml SKUs**, as indicated by the taller bars on the left side of the chart. This pattern suggests that smaller package sizes may be more attractive to consumers in the current context, potentially due to pricing, convenience, or usage patterns.

Figure 9



When we apply the same optimization process using **revenue** (Figure 9) as the objective function—instead of share of choice—the resulting portfolio landscape changes. While the two dominant SKUs, *"Yaddle"* and *"Moroff"*, continue to appear in all top configurations, the previously observed preference for 250ml SKUs is no longer evident. Interestingly, the revenue-based optimization now includes both the 250ml and **350ml variants** of the top-performing SKUs, suggesting that larger pack sizes may contribute more strongly to revenue, likely due to higher unit prices and margins.

By combining the **SKU profiles from top-performing solutions** with relevant **business constraints**, we can derive actionable guidelines for portfolio design. For instance, certain SKUs may emerge as **"must-haves"**—flagship products that are central to brand identity, brand heritage, or consumer recognition. Based on this, we can construct a **prioritized list** of additional SKUs to consider, forming a hierarchy that supports informed decisions about what to include next—especially in retailer-specific assortments.

The **Collaborative Optimization Framework** thus enables a nuanced understanding of portfolio potential. It integrates quantitative model outputs with qualitative business logic, supporting the creation of tailored, strategically sound assortments that go beyond purely mathematical optima.

Figure 10

You can find the optimal portfolio size. In Figure 11 you see that for the 250-ml SKUs 6, or even 5, SKUs can give you the same revenue as 8 300-ml SKUs.

Figure 11

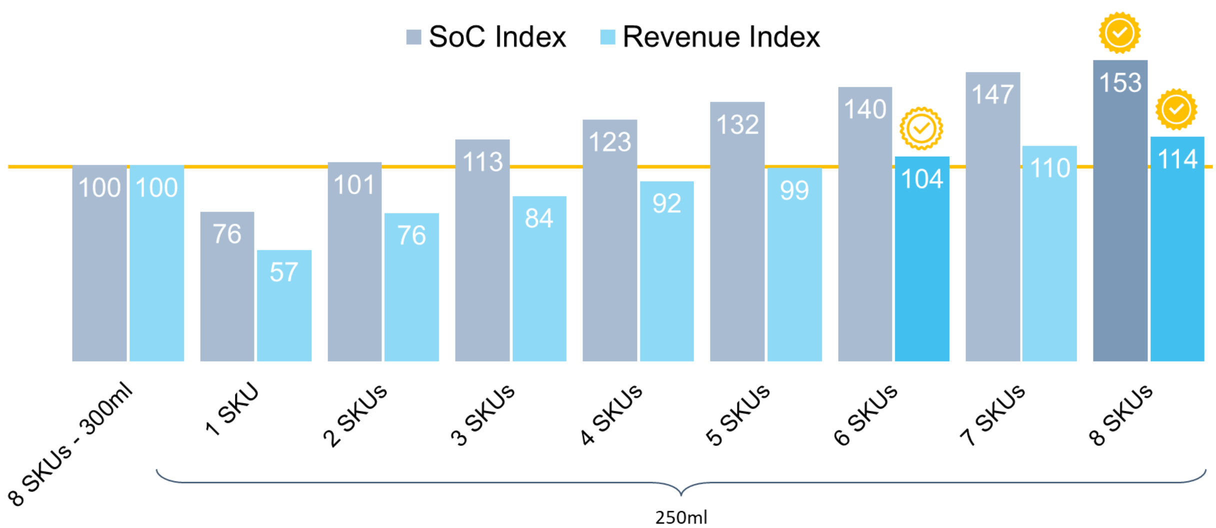
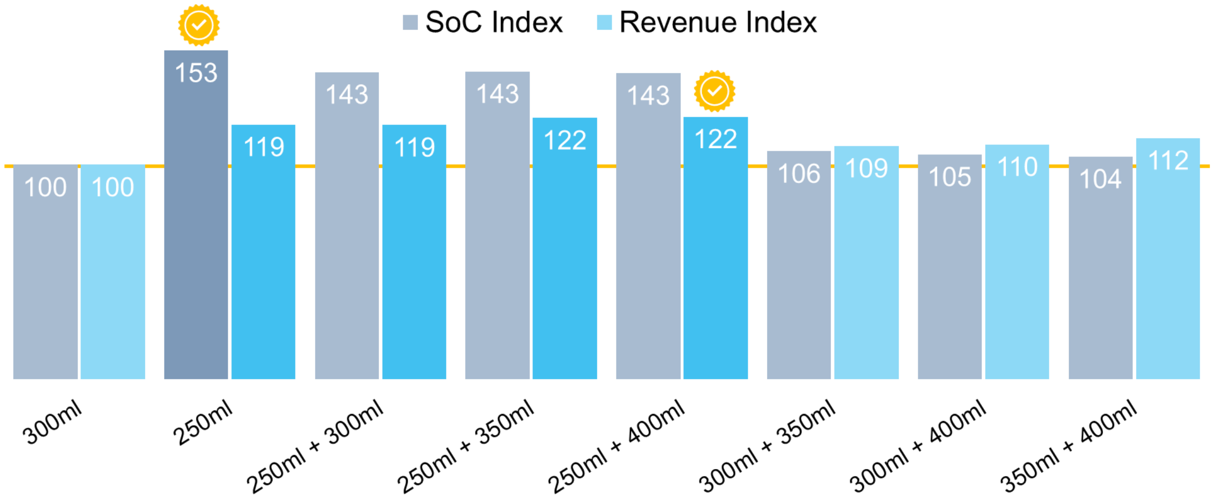


Figure 12 illustrates how the inclusion of multiple pack sizes—for example, combining 250ml and 400ml variants within the same total number of SKUs—can lead to higher revenue outcomes. This suggests that offering size variety, even without increasing the overall portfolio size, may better address different consumer needs and price sensitivities, thereby optimizing financial performance.

However, such configurations must be evaluated carefully from an **operational perspective**. Introducing additional pack sizes or maintaining a broader set of variants can lead to **higher production and logistics costs**. Therefore, any revenue gains must be weighed against potential increases in complexity and cost, highlighting the importance of integrating **cost data and operational feasibility** into the final portfolio decision-making process.

Figure 12



Comparison of Optimization Algorithms

To generate the results presented within the optimization framework, we implemented and tested **three different optimization approaches**: an **exhaustive search** (identifying the top 1,000 solutions), a **simulated annealing (SA)** algorithm, and a **genetic algorithm (GA)**. Each approach was evaluated based on two key criteria:

1. **Solution quality**—i.e., the ability to identify high-performing portfolios with respect to the defined objective functions, and
2. **Computational efficiency**—measured in terms of the time required to reach convergence or acceptable performance thresholds.

This comparative analysis allowed us to assess the **trade-offs** between accuracy and scalability. While the exhaustive approach serves as a benchmark by guaranteeing the global optimum (within the explored solution space), the heuristic methods (GA and SA) offer more practical alternatives for complex or high-dimensional problems where exhaustive search becomes computationally infeasible.

Figure 13

A comparison of **SKU inclusion frequencies** across the three optimization approaches (Figure 13) shows that all methods successfully identify the top-performing SKUs. However, the two heuristic algorithms—**Genetic Algorithm (GA)** and **Simulated Annealing (SA)**—display some **blind spots**, as certain SKUs do not appear in the resulting solution sets. In contrast, the **Exhaustive Search (Top 1000)** provides a more comprehensive picture of the solution landscape, offering greater transparency into the range and frequency of viable SKUs.

Despite this, GA and SA demonstrate a clear advantage in terms of **computational speed**. As illustrated in Figure 14, both heuristic methods converge to high-quality solutions significantly faster than the exhaustive approach. This efficiency, however, comes at the cost of a **more limited trace**—that is, fewer intermediate solutions are stored or explored during the optimization process, potentially reducing the depth of post-hoc analysis.

In summary, while the **exhaustive search** method provides a more comprehensive understanding of the solution space—capturing a wider range of high-performing configurations and SKU patterns—it is computationally demanding and often impractical for large-scale problems. In contrast, **genetic algorithms (GA)** and **simulated annealing (SA)** offer a **highly efficient means of identifying optimal or near-optimal portfolios**, delivering strong results with significantly reduced computation time. These heuristic approaches are particularly valuable in **time-sensitive contexts** or when working under **computational constraints**, making them practical tools for real-world portfolio design tasks.

Figure 14

Time in Minutes	SoC			Revenue		
	Top1000	**GA**	**SA**	**Top1000**	**GA**	**SA**
Simple	4,4*	2,9	1,2	4,3*	2,5	1,2
Intermediate	54,3	2,3	0,8	62,4	2,7	0,8
Complex	1359,3	29,3	26,8	1294,4	28,6	24,5

Focusing on the **complex example** discussed earlier, we observe an absolute contrast in computational effort across the optimization methods. The **Top 1,000 exhaustive search** requires nearly a full day of processing time, whereas the **heuristic algorithms**—particularly **Genetic Algorithm (GA)** and **Simulated Annealing (SA)**—complete their runs in approximately **30 minutes**.

This significant difference in runtime is explained by the **search behavior** of the algorithms. As illustrated in Figure 15, GA and SA reach optimal or near-optimal solutions **without exploring a large number of unique portfolio combinations**. Both algorithms are highly efficient in navigating the solution space and converging toward strong outcomes, especially GA, which demonstrates particularly fast convergence to the global optimum.

However, this efficiency comes with a trade-off: the **reduced diversity of explored solutions** provides **less analytical depth** for subsequent **profiling of top-performing portfolios**. In contrast, the exhaustive approach—despite being slower—yields a richer foundation for identifying patterns, understanding SKU relevance, and supporting strategic decision-making based on the frequency and structure of high-performing configurations.

Figure 15

SoC 8/32	time (min)	nCombs	Unique Combs	in Top1000
Top1000	1359,3	1000	1000	
GA	29,3	320	13	0,8
SA	26,8	4004	386	17,3

Given the strong performance of the **Genetic Algorithm (GA)**, we explored an approach to **accelerate the exhaustive Top 1,000 search** by using the GA-derived optimum as an **intelligent starting point**. Rather than beginning from a random portfolio and iteratively moving "up the hill" to identify high-performing combinations, this method **starts at the top**—the GA solution—and works **downward** to expand the base of high-quality portfolios for subsequent profiling.

For the exhaustive Top 1000 search we exchanged each SKU in the optimal solution by the SKUs that did not make the optimum. We saved these solutions and repeated the process for all resulting solutions (we kept the Top 1000 for each iteration). As soon as the Top 1000 did not change we stopped the process.

This hybrid approach yields a notable efficiency gain. As shown in Figure 16, computation time is reduced by a factor of approximately **2.5**. While the original exhaustive search required nearly 20 hours for the complex model, the GA-informed version completes in roughly **8 hours**—a duration that makes it **practically feasible as an overnight process**.

This result demonstrates the potential of **combining heuristic and exhaustive methods**: leveraging the speed and efficiency of GA to guide and streamline more computationally intensive profiling techniques.

Figure 16

	time (min)	Split
Top1000	62,4	Intermediate Rev
GA + Exhaustive	**28,7**	Intermediate Rev
Top1000	54,3	Intermediate SoC
GA + Exhaustive	**33,5**	Intermediate SoC
Top1000	1294,4	Complex Rev
GA + Exhaustive	**555,6**	Complex Rev
Top1000	1359,3	Complex SoC
GA + Exhaustive	**365,8**	Complex SoC

Average Improvement Factor
≈
2,5

Summary

In practice, we recommend the consideration of the following aspects (Figure 17):

Figure 17

- Exhaustive search gets you the **best combinations**
- All optimization algorithms derive the best solution and some “good” solutions
- Both GA and SA **find the best solution fast**. Especially GA navigates quite fast towards the optimum
- Combination of GA / SA and Top1000 search can speed up finding the best combinations for complex problems

Effective portfolio optimization requires a balance between computational rigor and strategic relevance. A practical approach begins with **exhaustive search**, where feasible, to establish a performance benchmark. By identifying the top 1,000 portfolio configurations, researchers gain a solid foundation for subsequent evaluation, including scenario comparisons and business-specific filtering.

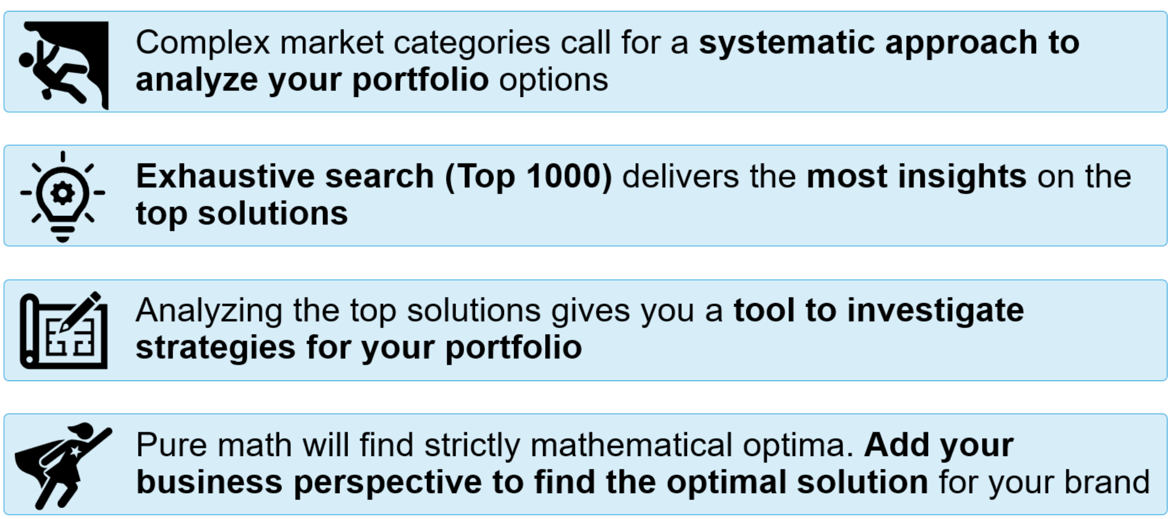
For more complex or high-dimensional problems, **heuristic methods** such as genetic algorithms (GA) or simulated annealing (SA) offer scalable alternatives. These algorithms can efficiently explore large solution spaces, where exhaustive enumeration would be computationally infeasible.

In practice, **hybrid strategies** can enhance solution quality. For example, a genetic algorithm may be used to broadly search the solution space, followed by a local optimization step—such as hill climbing or focused exhaustive search—to refine the best candidates.

Importantly, **quantitative optimization must always be complemented by managerial judgment**. Purely mathematical optima may fail to account for operational constraints, brand strategy, or consumer perceptions. A solution that maximizes utility or profit in the model may still be unfeasible in practice due to portfolio complexity, channel limitations, or cannibalization effects.

Therefore, successful portfolio design demands **close collaboration between data scientists and marketing professionals**. Only by combining analytical precision with strategic insight can organizations arrive at solutions that are not only optimal on paper, but also viable and effective in the market (Figure 18).

Figure 18



Maximilian Rausch

Peter Kurz

Stefan Binner

References

Balakrishnan, P.S.; Jakob, V,S, (1996): Genetic Algorithms for Product Design, in: Management Science, 42 (8), 1105–1117.

Bayer, D.; Voekler, S. (2023): One-stage product-line design heuristics: an empirical comparison. OR Spectrum 46, 73–107.

Belloni A.; Freund, R.; Selove, M.; Simester, D. (2008): Optimizing Product Line Designs: Efficient Methods and Comparisons, in: Management Science, 54 (9), 1544–1552.

Holland, J.H. (1975): Adaption in natural and artificial systems; An introductory analysis with applications to biology, control and artificial intelligence., Oxford, UK: University of Michigan Press.

Kirkpattrik, S.; Gelatt Jr, C.D.; Vecchi, M.P. (1983): Optimization by Simulated Annealing, in: Science, 220 (4598), 671–680.

McFadden, D. (1976): Quantal choice analysis: a survey. Ann Econ Soc Meas 5:363–90.

Selzam, B. (2023); Genetische Algorithmen (Ausarbeitung) Projektgruppe 431 TU Dortmind (Working Paper).

Tsafarakis, S.; Marinakis, Y.; Matsatsinis, N. (2011): Particle swarm optimization for optimal product line design, in: International Journal of Research in Marketing, 28 (1), 13–22.

TOWARD A SMARTER MAXDIFF: RETHINKING SOME CONVENTIONAL STRATEGIES

MING SHAN
HALL & PARTNERS

ABSTRACT

A typical MaxDiff is carried out sequentially: design precedes data collection, and modeling comes last. Inspired by Sawtooth's Adaptive and Bandit MaxDiff approaches and the concept of computerized adaptive testing, this research explores methods to model each individual respondent during the MaxDiff survey, along with any already completed sub-sample, to inform design in real time for performance gains. Item Response Theory (IRT) is chosen as the primary modeling tool for its speed and rich diagnostics. Simulations show that such a "smarter" MaxDiff can outperform standard MaxDiff designs. This also suggests that a balanced design is often suboptimal, and that MaxDiff design should be data- and goal-dependent. To lower the barrier to implementing such dynamic designs, I propose a simple user-specification input mechanism.

INTRODUCTION

MaxDiff (Louviere, 1991) is a widely used and powerful technique. The traditional implementation of MaxDiff can be roughly characterized by the following steps:

- Create a balanced design based on the number of items.
- All respondents complete the same number of tasks from a randomly assigned version.
- The model is estimated after data collection, typically using Hierarchical Bayes (HB).

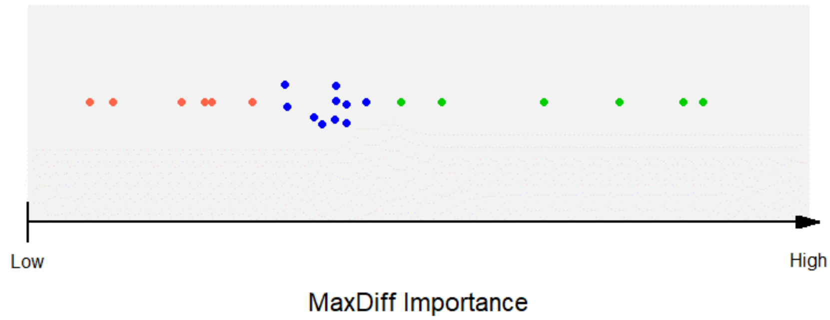
These steps are sequential. Express and Sparse MaxDiff, designed to address larger numbers of items, follow this structure. An exception is Sawtooth's Adaptive and Bandit MaxDiff (Orme, 2006; Fairchild et al., 2015), which learn from a subset of respondents to generate customized task designs mid-survey, focusing on items of greater importance.

This research examines areas where current MaxDiff practices may fall short and explores general paths for improvement. The paper begins by introducing the problem, followed by a brief overview of Item Response Theory and Computerized Adaptive Testing. The idea of a smarter MaxDiff is then discussed, supported by simulations and concluding remarks.

RESEARCH PROBLEM

Figure 1 shows the HB-estimated importance from a real MaxDiff study of 22 items, using a standard design with 100 versions, 5 items per task, and 13 tasks per respondent. If we divide the importance into three tiers, 6 high (H) items, 10 middle (M) items, and 6 low (L) items, a striking pattern emerges: the 10 middle items (represented as blue dots) are crowded within a very narrow range. Some are nearly indistinguishable in this case, and their positions are jittered in the plot to improve visual clarity.

Figure 1: Average MaxDiff Importance from a Real Study (22 Items)



The distribution of tier combinations across all 1,300 tasks is shown in Table 1. It reveals that 62% of MaxDiff tasks include items from all three tiers. Such tasks are likely too easy for respondents and provide limited information about the middle-tier items, as these are often neither selected as the most nor the least important.

Table 1: Distribution of Tasks by Tier Combinations

HL	HM	HML	M	ML
3%	17%	62%	1%	17%

This raises the question: should MaxDiff design be more tailored to the respondents and study objectives, rather than strictly balanced and predetermined, as is traditionally done? More broadly, are there further opportunities to make MaxDiff perform better?

My goal is to rethink some of the conventional strategies and explore alternatives.

Some of my hypotheses:

- A pre-made and balanced MaxDiff design may not be the ideal option—at least in this case.
- It may be possible to make MaxDiff smarter and adaptive to the data.
- If so, how can such an adaptive approach be generalized for broader implementation?

Strategies

Among the many possible areas for improvement throughout the MaxDiff process, this research focuses primarily on real-time design and modeling through individualized learning within each MaxDiff interview.

To support this idea, I briefly discuss two well-established topics: **Item Response Theory (IRT)** and **Computerized Adaptive Testing (CAT)**. Both have been widely researched and applied, and they help frame the ideas proposed in this paper.

This search for a smarter MaxDiff centers on:

- Modeling across—and even within—respondents to guide design during data collection.
- Allowing the design to be driven by a combination of user specifications and ongoing learning.

I rely on simulations to illustrate the effectiveness of the proposed approach.

ITEM RESPONSE THEORY (IRT)

IRT is a modern psychometric approach for test design, modeling, and scoring (e.g., de Ayala, 2022; van der Linden, 2018), which is also referred to as modern test theory. The input data for a typical IRT model is in a rectangular form of N respondents by Q items (or answers). The value for each item in Q can be binary, categorical, or continuous, but most are coded dichotomously as 0 and 1 to indicate if an answer is correct or wrong. IRT has many different model forms and can be linked to many commonly applied statistical models at its root (Chen et al., 2021). Rating scale and partial credit models can accommodate polytomous responses instead of binary coded data. Certain IRT models allow a specific parameter to capture the degree of guessing by respondents. IRT is a tool to learn and capture the key characteristics of both respondents and items. The key IRT model outputs include the ability scores of individual respondents (as the overall score from a test) and a difficulty score of each item (to help understand the quality or inclusion of a test question). Item difficulty and respondent ability are placed on the *same continuous scale*. For this proposed research, the plan is to restructure and then feed MaxDiff answers into a proper IRT model and use the estimated item difficulty scores to equate to the importance scores from MaxDiff.

There are a variety of IRT models dealing with different response types including but not limited to Rasch and 2–4 Parameter Logistic (PL) for binary values, Graded Item Response Model for ordinal responses, and Continuous Response Model for continuous values. After exploration and comparison, I found Rasch (1PL) and 2PL models, two simpler forms, work quite well for this particular purpose. With a very limited number of parameters, estimation runs very fast and always converges.

Two-Parameter Logistic Model (2PL)

Equation 1 is the formula for the 2PL model. I use lowercase *p* to denote a person and lowercase *i* to denote an item. Again, an item refers to a test question a respondent answers. We model the probability of respondent *p* answering item *i* correctly, which is a raw binary input variable in this case. Theta (θ) is person-specific and measures their ability, similar to a final test score. The two parameters, *a* and *b*, are tied to items: *b* is the intercept that dictates where the logistic curve turns on the horizontal scale and measures item difficulty; *a* is the slope and measures discrimination. Equation 2 is a simplified version of the 2PL model and closely resembles a logistic regression model. In terms of a test, an easy item means more test takers will answer it correctly, while a more discriminating item better separates respondents by ability.

$$\boldsymbol{Prob}\left(\boldsymbol{y_{pi}} = \mathbf{1} \middle| \boldsymbol{\theta_p}, \boldsymbol{a_i}, \boldsymbol{b_i}\right) = \frac{\boldsymbol{exp}\left(\boldsymbol{a_i}\left(\boldsymbol{\theta_p} - \boldsymbol{b_i}\right)\right)}{\mathbf{1} + \boldsymbol{exp}\left(\boldsymbol{a_i}\left(\boldsymbol{\theta_p} - \boldsymbol{b_i}\right)\right)} = \frac{\mathbf{1}}{\mathbf{1} + \boldsymbol{exp}\left(-\boldsymbol{a_i}\left(\boldsymbol{\theta_p} - \boldsymbol{b_i}\right)\right)} \quad (1)$$

$$\boldsymbol{log}(\boldsymbol{P}/(\mathbf{1} - \boldsymbol{P})) = \boldsymbol{a_i}\left(\boldsymbol{\theta_p} - \boldsymbol{b_i}\right) \quad (2)$$

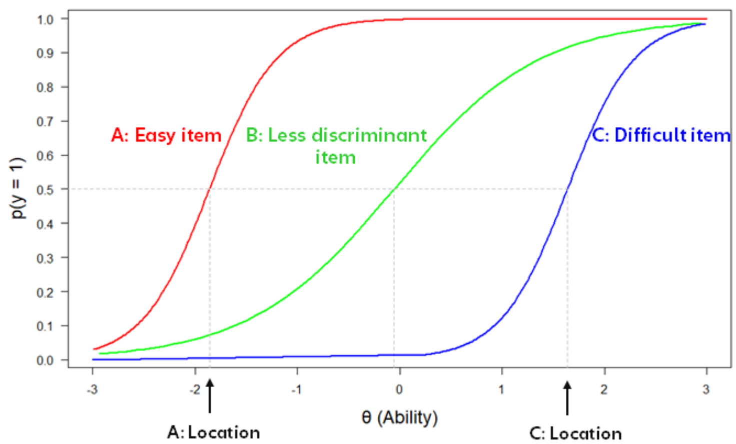
Where:

- y_{pi} is the event that person p answers item i correctly.
- θ_p is the ability parameter unique to person p, akin to a single test score.
- b_i is the difficulty parameter unique to item $\boldsymbol{i}$.
- a_i is the discrimination parameter unique to item i.

Item Response Function (IRF)—An Illustration of 3 Items

An example shown in Figure 2 offers some further clarification. Using Equation 1, we can plot the probability for each item over the entire range of θ, which represents the respondent's ability level. Such a curve is called the *item response function*. For each curve, we can find the θ value that corresponds to a 50% probability—that is, a half chance of getting the answer right and half the chance of getting it wrong on that question. This specific θ value is identified as the location of the item. At this midpoint, it is also the inflection point of the logistic curve, as shown by the dotted gray lines indicating the location. An item located further left is easier, while an item located further right is more difficult. For example, at the value -2 on the x-axis, about 50% of test takers get item A correct, fewer than 10% get item B correct, and almost nobody gets item C correct. The middle item in green is flatter, meaning it has lower discrimination, since it does not have a very distinct θ point to separate high and low probabilities. To use the IRT model to estimate MaxDiff data, I chose item locations to represent item importance.

Figure 2: Item Response Function (IRF)



Using 2PL Models to Estimate MaxDiff Importance

Table 2 shows a typical MaxDiff data structure, presenting just the first two tasks for a respondent. Table 3 shows how I code the MaxDiff response data into IRT model input. To fit a logistic outcome, each task is expanded into two rows: one row for the best choice versus the rest, and the second row for the worst choice versus the rest.

From the first row of Table 2, we see item 19 is picked as the best and item 10 as the worst out of the 4 items: 1, 7, 10, and 19. In Table 3, the first row is for the best-versus-others comparison: item 19 is coded as 1, and items 1, 7, and 10—being in the same task set—are coded as 0. The second row is for the worst-versus-others comparison: item 10, being the worst, is coded as 0, and items 1, 7, and 19 are coded as 1. Items not included in the current task are left as missing.

Table 2: MaxDiff Data—Original Structure

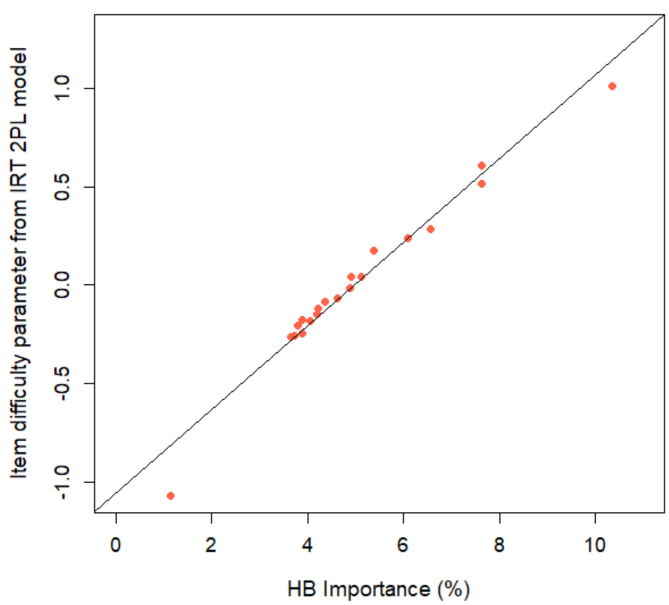
respondent	version	task	item1	item2	item3	item4	best.pick	worst.pick
resp1	1	1	10	1	19	7	19	10
resp1	1	2	19	17	15	3	3	15
...	...	...	...	...	...	...	...	...

Table 3: MaxDiff Data—Recoded for 2PL IRT Model

resp.id	task	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10	i11	i12	i13	i14	i15	i16	i17	i18	i19	i20
resp1	1 (best vs. rest)	0	.	.	.	.	.	0	.	.	0	.	.	.	.	.	.	.	.	1	.
resp1	1 (rest vs. worst)	1	.	.	.	.	.	1	.	.	0	.	.	.	.	.	.	.	.	1	.
resp1	2 (best vs. rest)	.	.	1	.	.	.	.	.	.	.	.	.	.	.	0	.	0	.	0	.
resp1	2 (rest vs. worst)	.	.	1	.	.	.	.	.	.	.	.	.	.	.	0	.	1	.	1	.
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Figure 3 compares the MaxDiff importance results from the 2PL IRT model and the HB model on Sawtooth's fictitious Ski example data. The dataset includes 250 respondents, 20 items, and 20 tasks per respondent. The correlation between the two sets of estimated importance scores is 0.98. It should be emphasized that IRT is used for quick and interim response pattern learning, while HB is still used for the *final model estimation*.

Figure 3: HB Estimate vs. IRT Item Difficulty Parameter



Computerized Adaptive Testing

In the testing world, many large-scale tests (e.g., student academic assessments) that used to be carried out in the "one-single-test-for-all" form have been replaced by computerized adaptive testing (e.g., Weiss and Sahin, 2024), or CAT. CAT learns a test taker's ability and scores each question upon its completion during the test. The algorithm updates the person's ability score, then uses it to determine the assignment of the next question and when to end the test based on the most updated learning about that individual.

Among the many benefits of the CAT approach, the following are worth mentioning:

- **Individually targeted test:** Assigns test questions matching a test taker's ability to maximize information extraction and avoid giving overly easy questions (which result in most answers being correct) or overly difficult questions (which may force guessing). Both situations reduce discrimination.
- **Fewer test questions** are generally required.
- **The number of questions** does not need to be the same across test takers. Achieved accuracy dictates how many tasks are needed.

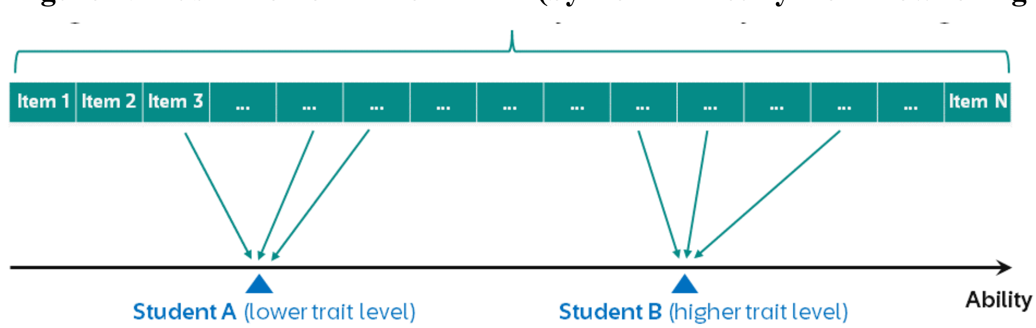
Both Adaptive/Bandit MaxDiff from Sawtooth and CAT share some similarities. They all learn and use data patterns during the data collection to inform and adjust the design. While Sawtooth's approaches use the early part of the sample to approximate the final aggregated patterns and help adjust the design for the remaining survey sample (by way of cross-respondents), CAT focuses on within-individual learning and adaptive questioning.

The CAT idea can be viewed as a framework for addressing broader and more generic situations. In the example given earlier, the focus required could well be on the accuracy of the middle (rather than the most important) items. Challenges—such as the requirement in CAT for a pre-built test question bank, while each MaxDiff study is typically brand new—need to be circumvented, or approximations made, in order to implement the general ideas of within-individual learning and the adaptive nature of CAT design.

CAT relies on IRT as its theoretical foundation. As mentioned earlier, IRT can place item difficulties and examinees' abilities on a single continuous scale. CAT takes an iterative, adaptive process to learn a person's ability during the exam and help decide the next best test question to give. The intention is to assign test items that best match one's ability level.

There are different criteria for terminating a test. It can end when a pre-set precision target is met, when the maximum number of questions or time limit is reached, or based on other conditions. In the hypothetical example illustrated by Figure 4, once it is learned that Student A has lower ability, items of lower difficulty are assigned. This contrasts with Student B, who has higher ability and will receive a series of more difficult items.

Figure 4: Illustration of an Item Bank (by Item Difficulty from Low to High)



It should be noted that in CAT, the item bank needs to be carefully constructed and tested to meet certain desired properties. For the MaxDiff problem, the focus is not on the individual but rather on group-level importance scores and rankings. Therefore, we can afford to loosen the requirements and instead borrow the concept. The “bank” is the item set for each study. We update the bank as more respondents complete the survey.

Figure 5 shows the flow for a single person being tested. For the very first test question, either an item of average difficulty is given, or—ideally—prior knowledge about the person is used to assign an item that best matches their estimated difficulty level. After each new answer is provided, the person’s ability estimate is updated. If the stopping rule has not yet been met, a new question will be selected from the item bank to continue the test. If the stopping rule is met, the test ends and a final score is calculated.

This intertwined process of learning ability and assigning matching test questions leads to a shorter test and more accurate scoring.

Figure 5: CAT Flow for a Single Person

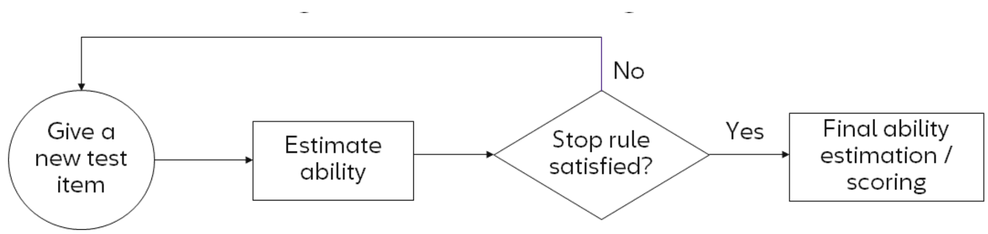


Figure 6a–6b is generated using the R package *catR* (Magis and Raîche, 2012) to illustrate the process for two different individuals. Using Person A as an example: with just the first test item, the accuracy range is very wide. After a correct answer, CAT assigns a second test item of higher difficulty. Another correct answer leads to an even more difficult question. An incorrect response on the fourth question results in a less difficult item being administered. This process continues.

As the test progresses, the remaining items and ability estimates become more closely aligned, leading to a shrinkage in the measurement error band. The test terminates after 15 items, once the error range falls below the targeted threshold.

Figure 6a: Person A Tested via 15 Items

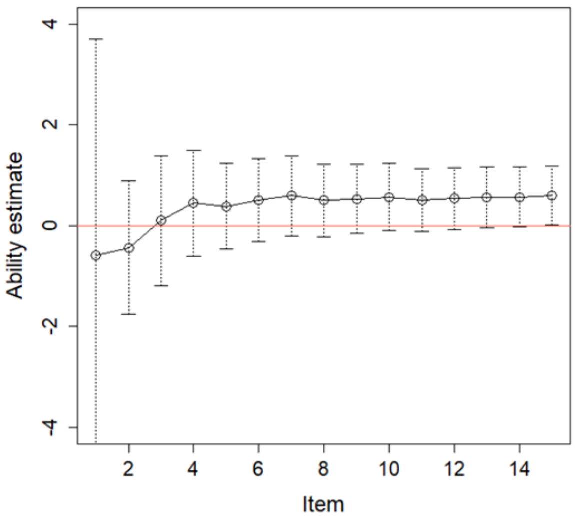


Figure 6b: Person B Tested via 18 Items

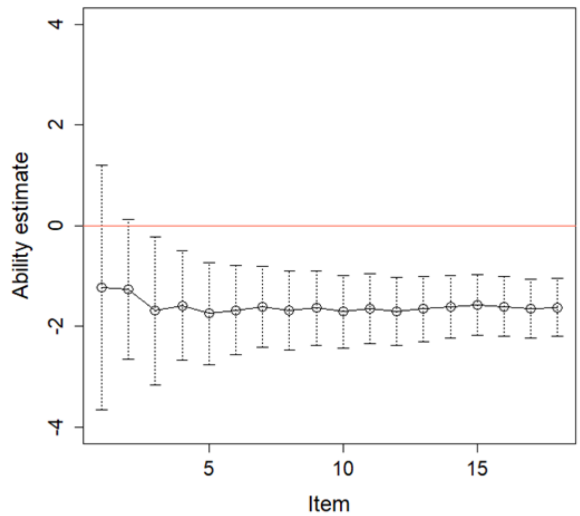
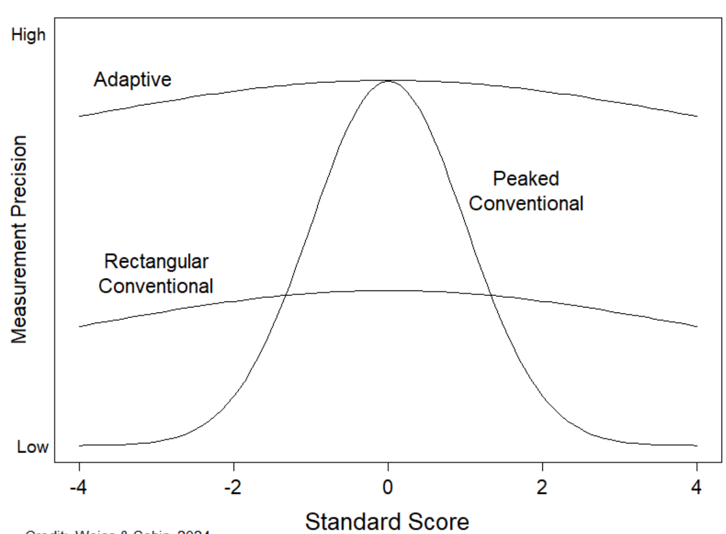


Figure 7 is a replication of Figure 1-1 on page 5 of Weiss and Sahin (2024), demonstrating the benefit of adaptive testing. Both the peaked and rectangular tests are conventional formats. The peaked test, with its bell-shaped curve, concentrates test items around average difficulty. As a result, it provides higher accuracy for respondents with average ability, but not for those with very low or high abilities. Low-ability respondents will get most questions wrong, while high-ability respondents will get most questions right.

The rectangular test spreads item difficulty more evenly across the range. However, this approach is also less efficient, as some test items will be either too difficult or too easy for all respondents. This is similar to the conventional balanced MaxDiff design.

In contrast, due to the better match between test items and respondents' trait levels, the adaptive test provides higher accuracy across the full range of abilities.

Figure 7: Illustration of Different Measurement Precisions



Credit: Weiss & Sahin, 2024

MAKE STANDARD MAXDIFF SMARTER?

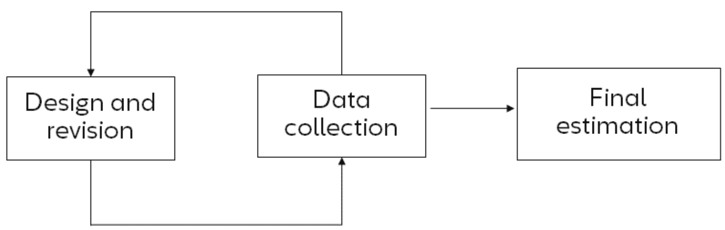
Figure 8 highlights again the sequential nature of a standard MaxDiff: a balanced design, but without any learning.

Figure 8: Sequential and Fixed Nature of Standard MaxDiff

Two approaches can be taken to learn from data already collected to improve the design: *across* respondents and *within* each respondent. For learning across respondents, we can have the analysis from any completed respondents inform the design for the following respondents. For example, an item bank can be constructed, or the item locations can be improved. The emerging importance pattern helps determine whether certain items should be over- or under-exposed.

For learning within each respondent, we can use the first *m* tasks completed to inform what to show in the remaining tasks. For example, we can estimate the respondent's ability or compute data fit statistics and use them to guide whether easier or more difficult tasks should be assigned. Like CAT, we could even vary the total number of tasks per respondent or display different numbers of tasks across tasks. This general flow of learning from already collected data is illustrated in Figure 9.

Figure 9: Learn from Already Collected Data to Improve the Design



For any model to be practical during the data collection phase, it needs to run in just a split second to avoid any survey delay, and it must be able to handle a very small amount of data. IRT and aggregate logit models can be considered, but not HB due to its longer estimation time. One of the appealing factors in choosing IRT over aggregate logit is its diagnostic outputs (e.g., item discrimination, the guessing parameter when explicitly specified, and person fit). Also, IRT models can be easily estimated (e.g., Chalmers, 2012) with just a handful of respondents, or even within a single respondent. For IRT models on a single completed task, I applied a trick by including extra artificial task(s) assigned with a very low weight. This avoids model failure but still ensures that only the completed task is reflected in the model output.

There are different levers I used to inform how the design can be modified. First, I use a simple one-dimensional array of length equal to the total number of items to specify how many times to over- or under-expose each item relative to the average number of times shown in a standard balanced design. An example will be given in the simulation section next. This array can be modified by users according to their own preferences or the specific study focus. The default setting is recommended to be in ascending order by item importance, giving more exposure to items of higher importance. The program also generates a recommendation based on the density of item distribution on importance and their variances. The modification array can also be specified as multipliers, and it will be converted into absolute adjusters by the program to meet the total number of exposures allowed. When the adjusters are fractional, the program allows slightly different exposure frequencies across respondents to meet the overall target. Any of the levers for design adjustment can be turned off. If all are turned off, the standard MaxDiff design is applied.

Here are some additional details on how the design is modified to meet the adjustment specifications:

- Always start from a standard MaxDiff fixed design with plenty of extra tasks per version.
- When a respondent's assigned version needs adjustment, the item with the highest importance has priority.
- When deciding on a replacement item, one that is closest in importance to the rest of the set is first selected for a tougher trade-off.
- Choose modifications that result in the lowest correlation (for design efficiency).
- Each modification step is based on a weighted consideration of these different factors.

Additionally, some safeguards are also put in place in my experiment. For example, after a search, if no modified design can meet all pre-set requirements, a warning is given. A simulation can also be run ahead of time to get a sense of the potential impact.

It should be noted that the above setup is for exploration and proof of concept purposes, and it is not yet optimal. As of now, the importance density- or variance-based information feeding into the design still requires more development. Also, the design modification is mostly driven by across-respondent learning rather than within-respondent learning.

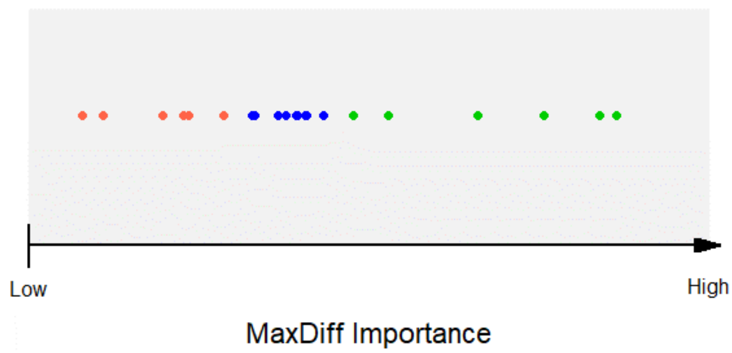
SIMULATION

The main hypothesis of this paper is that a smarter MaxDiff, with a design adaptive to the data, performs better than a conventional MaxDiff. The former is assigned as the treatment group and the latter as the control group. Simulation is a good way to make comparisons between the two groups while controlling for other conditions. This section discusses some of the simulation results.

Simulation Specifications

I try to mimic the importance pattern of the example introduced at the very beginning of this paper. Since that is a model-estimated result, I chose not to use it directly as the assumption to avoid any potential bias. The importance distribution shown in Figure 10 is used as the population truth. It is generated by a truncated normal distribution with the lowest possible value of 0.

Figure 10: Hypothetical Average MaxDiff Importance (22 Items)



Each respondent goes through 13 tasks and sees 5 items per screen, with a randomly assigned version. For every respondent, I randomly vary each importance value from the population truth using a normally distributed error with standard deviation equal to the importance itself. So, larger importance values will have higher variation. For each task, the best and worst choices are determined by the magnitude of the importance values after randomization.

I tested sample sizes from 150 to 500 in increments of 50. At each sample size, 10 separate sample draws were made. For each draw, I simulated two scenarios: a standard MaxDiff with no design modification to serve as the control scenario, and a user-specified modification array applied to serve as the treatment scenario. To isolate the effect, only the specified modification takes effect—nothing else.

Simulation A—Over Sample Increasingly by Item Importance

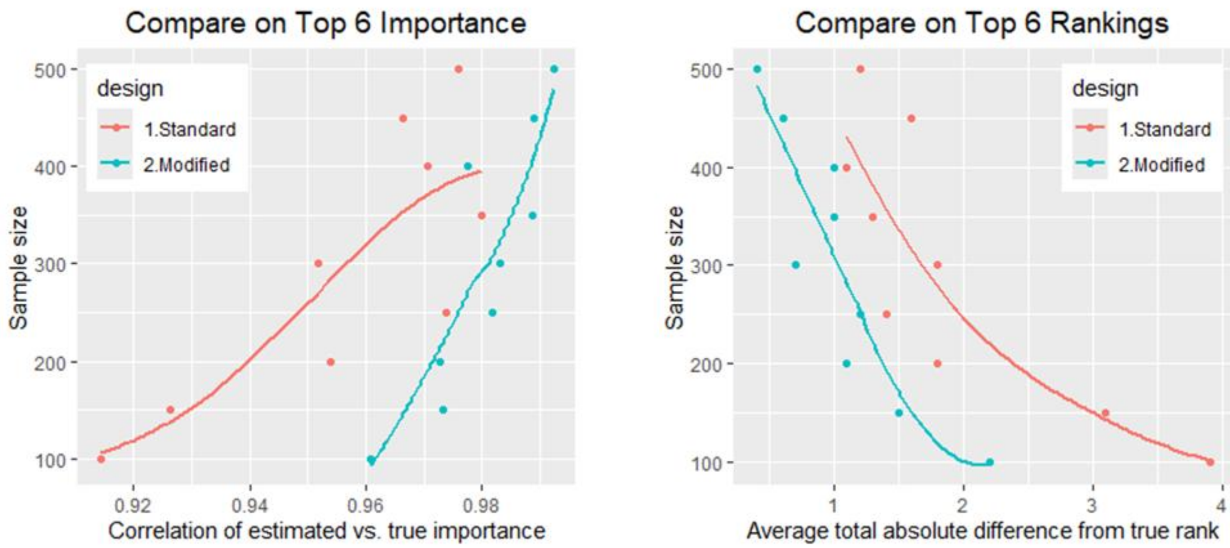
Table 4 gives the modification array for Simulation A. For simplicity and demonstration purposes, item exposures in the design were only modified by integers. Without modification, each item was seen about 3 times by each respondent. This increased exposure by importance could be reasonably applied to typical MaxDiff studies where the focus is on the top items.

Table 4: Item Exposure Adjustment Array Increasing by Item Importance

Lowest importance																					*Highest importance*
i1	i2	i3	i4	i5	i6	i7	i8	i9	i10	i11	i12	i13	i14	i15	i16	i17	i18	i19	i20	i21	i22
-2	-2	-2	-2	-2	-2	-1	-1	-1	-1	-1	1	1	1	1	1	2	2	2	2	2	2

Figure 11 compares the total MaxDiff importance from the two simulated groups: one from the standard and balanced MaxDiff design, and another with the modified design per Table 4. The left chart compares the correlation of estimated importance with the importance truth, which we know. The chart on the right compares the absolute difference in importance ranking from the truth. Each dot is the average of the 10 draws at a given sample size. The modified design had a higher correlation and lower ranking difference, meaning the importance it recovered was closer to the truth, so it outperformed the standard MaxDiff on these two measurements. The total sample importance was the average of the individual respondent importance. Even with 10 draws, we saw quite a bit of fluctuation. The non-linear curves gave a feel about the general pattern. The vertical gap between the two lines represented the amount of extra sample required to achieve the same precision.

Figure 11: Simulation A Result—Standard vs. Modified Design on Top 6 Importance



Simulation B—Over Sample Items of Mid-level Importance

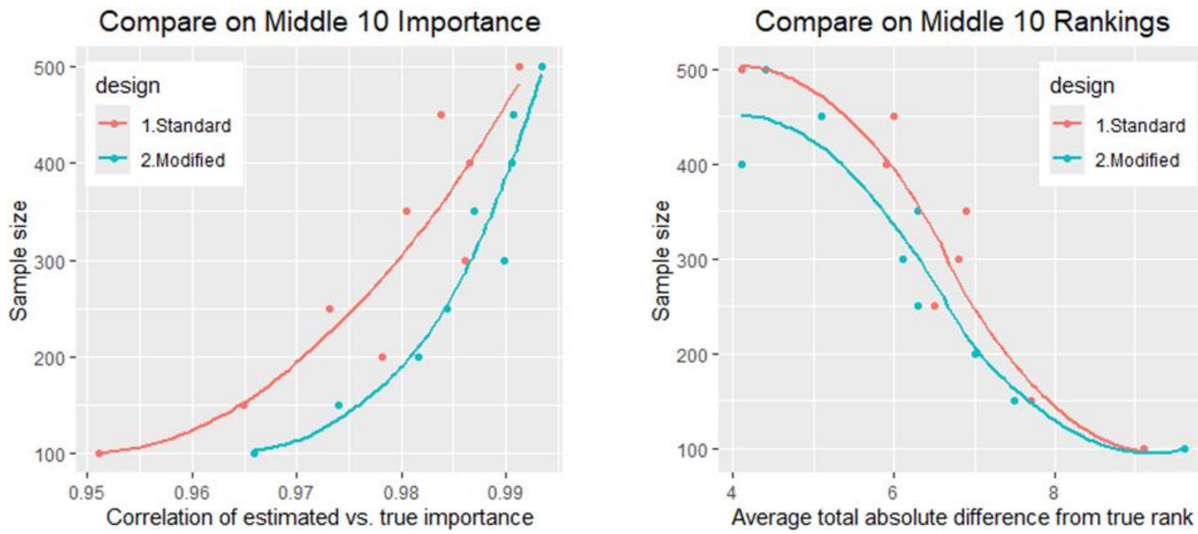
Table 5 gives the modification specification for Simulation B. Items 7 to 16, the 10 middle-tier items, were overexposed. This was achieved by underexposing the bottom 5 items. These overexposures came from the underexposures of the 5 items of the lowest importance, while the top 6 items were left unadjusted.

Table 5: Item Exposure Adjustment Array that Over Samples the Middle

Lowest importance																					*Highest importance*
i1	**i2**	**i3**	**i4**	**i5**	**i6**	**i7**	**i8**	**i9**	**i10**	**i11**	**i12**	**i13**	**i14**	**i15**	**i16**	**i17**	**i18**	**i19**	**i20**	**i21**	**i22**
-2	-2	-2	-2	-2	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0

Figure 12 is a comparison like what we just saw for Simulation A. The difference is that we compare the middle 10 items that were overexposed. Again, the importance was averaged first, and the correlation and rank difference were calculated on these averaged sample totals. Although the modified group still outperformed the standard MaxDiff, the gaps were smaller, suggesting the precision gain was harder to achieve for the mid-level items.

Figure 12: Simulation B Result—Standard vs. Modified Design on Middle 10 Importance



Still from Simulation B, it was somewhat a surprise to see the pattern in Figure 13, which compares the top 6 item importance. Although no overexposure on the top 6 items was made, this shows some improvement in their precision. This could be partly due to the increased trade-offs between top- and mid-tier items.

Figure 13: Simulation B Result—Standard vs. Modified Design on Top 6 Importance

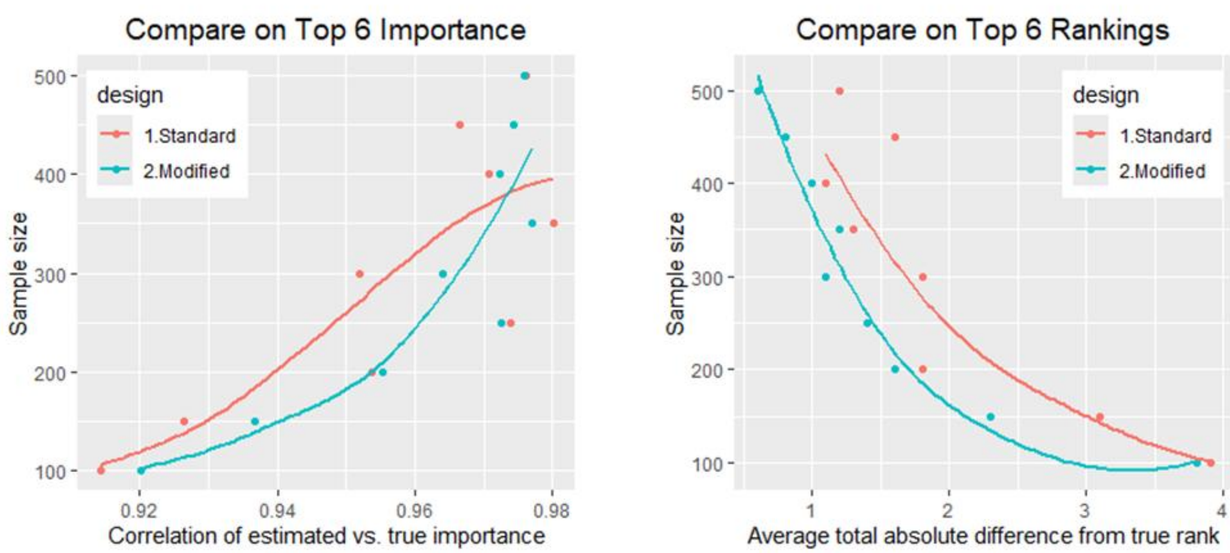
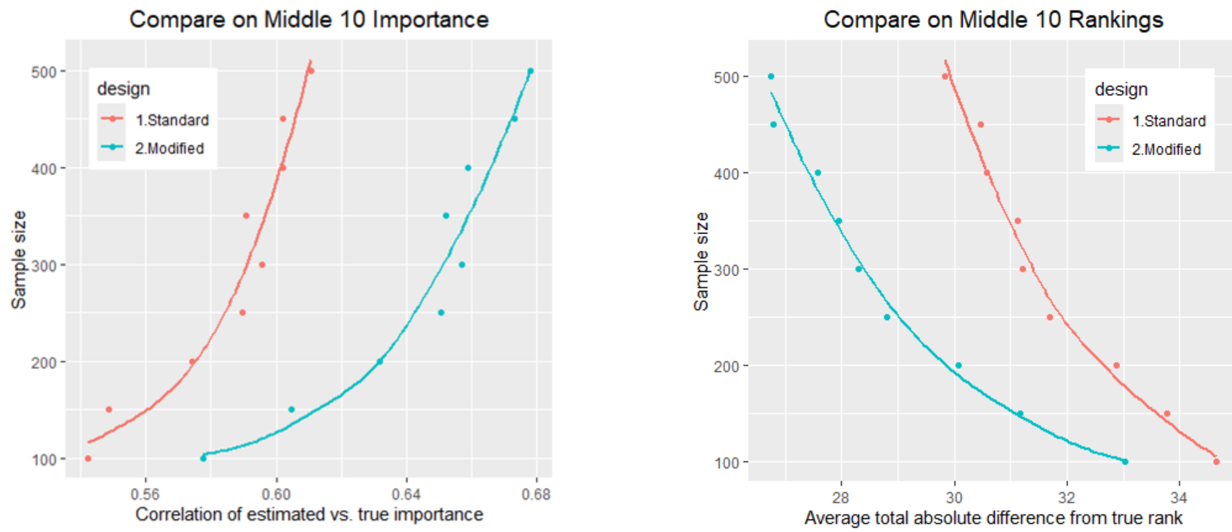


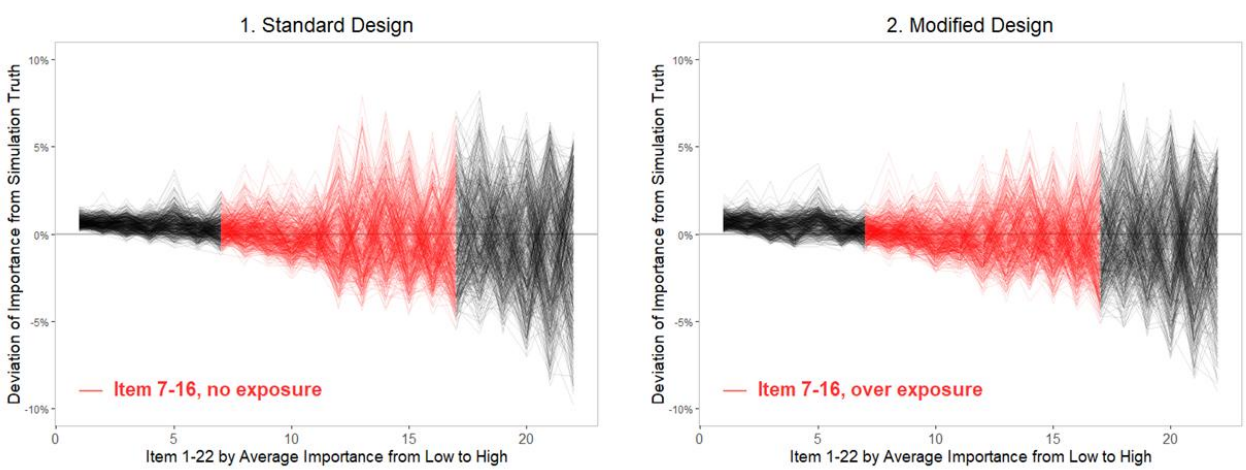
Figure 14 is still from Simulation B, but I focused on the respondent-level comparison in contrast to the total-level comparison we have seen so far. The correlation and absolute rank difference were calculated using the respondent-level importance estimates from HB for each *individual respondent first*, and then they were averaged to the sample total. We see much less sampling variation and more pronounced precision gains at the individual respondent level.

Figure 14: Simulation B Result—Standard vs. Modified Design on Middle 10 Importance



Lastly, Figure 15 further zooms in on individuals. I picked the very last of the 10 draws of Simulation B at the sample size of 500. Each chart plots all individual importance patterns from low to high—in this case, 500 lines for 500 respondents. I chose to show the differences from the truth instead of the absolute importance values. Items 7–16 in the middle were highlighted in red. The red section for the modified design in the second chart appeared tighter than the control group on the left. For those interested in using respondent-level importance for more in-depth analysis like segmentation, better precision at the respondent level would be welcome news.

Figure 15: Simulation B Result—Standard vs. Modified Design on 500 Individual Importance Patterns



Here were a few key takeaways from the simulations:

- Accuracy improved by item over-exposure, or the same precision was achieved with less sample.
- Precision gain outweighed the efficiency loss.
- This was particularly worth considering over standard MaxDiff when the importance patterns were unconventional, or the research focus was not solely on the most important items.
- Noticeable respondent-level accuracy gains were observed.

CONCLUSION

Before summarizing the general conclusion, I would like to point out some similarities and particularly differences between the current research and Bandit MaxDiff from Sawtooth. Both apply unbalanced designs and are applicable for the entire importance range. Bandit MaxDiff tends to focus on large item sets and the most important items. The focus of my investigation is not on large item sets to identify the most important items. At least in the specific problem shown, I looked at a moderate number of items, especially those mid-tier importance items that lack separation. Bandit MaxDiff has been shown to perform very well in addressing some challenges facing a standard MaxDiff (Fairchild et al., 2015). Using a few ideas from IRT and CAT, this research explored a somewhat different path. The goal was not to build a full solution to compare performance against Bandit MaxDiff. Instead, this research was set to dive further into across-respondent learning and especially within-respondent learning. I believe the within-respondent learning, i.e., a micro-view of respondents during the survey, is more promising than my work shows.

My intention was to draw attention to a few general issues. First, perhaps it is safe to say that the vast majority of MaxDiff studies are carried out on a moderate number of items following the standard approach. This design-first approach completely ignores the underlying importance pattern of each unique item set under study. An argument is made and some evidence is shown that data-dependent (e.g., importance distribution) and goal-dependent (e.g., picking top few winners) design—unbalanced and adaptive—can be more efficient and accurate. If one only cares about some top items, and I speculate many studies do, as shown by the simulations in this paper and argued elsewhere, that goal can be achieved with significantly less sample by just over-exposing the top items. If this is so obvious, a question then becomes: why shouldn't unbalanced design always be used? This leads to the second general issue I set to explore and my conclusion. For a data-dependent and smarter MaxDiff to work, there must be a good system able to learn from the data, allow limited human intervention, and perform reliably in different situations. Any effort to lower these barriers would be worthwhile. The design adjustment array and running simulations as a pre-check proposed here are just a couple of examples.

More work can certainly be done applying the across- and within-respondent learning framework. Some ideas mentioned next were explored but intentionally left out to keep my messages focused. A math-based adaptive design through some optimization algorithm would be preferred over the heuristic design modification strategy I applied. More respondent-level learning should be used to guide design adjustment. Multiple dimensionalities of MaxDiff importance patterns (i.e., multiple segments) certainly add complexity to the problems but need to be accounted for to accurately adapt to the underlying pattern and produce better and more valid results for real studies. Generalization of any approach for large item sets, which have unique challenges, is also necessary. Lastly, some empirical testing and comparison will help further validate and refine these approaches.

Ming Shan

REFERENCES

Chalmers, R. P. (2012). “mirt: A Multidimensional Item Response Theory Package for the R Environment,” Journal of Statistical Software, 48(6), 1–29.

Y. Chen, X. Li, J. Liu, and Z. Ying (2021), “Item Response Theory—A Statistical Framework for Educational and Psychological Measurement,” arXiv:2108.08604.

de Ayala, R. J. (2022), The Theory and Practice of Item Response Theory, 2nd edition, Guilford Press.

Fairchild, Kenneth, Bryan Orme, and Eric Schwartz (2015), “Bandit Adaptive MaxDiff for Huge Number of Items,” 2015 Sawtooth Software Conference, Provo, UT.

Louviere, J. J. (1991), “Best-Worst Scaling: A Model for the Largest Difference Judgments,” Working Paper, University of Alberta.

Magis, D., & Raîche, G. (2012). “Random Generation of Response Patterns under Computerized Adaptive Testing with the R Package catR,” Journal of Statistical Software, 48(8), 1–31.

Orme, Bryan (2006), “Adaptive Maximum Difference Scaling,” Sawtooth Software Research Paper.

van der Linden, W. J. (Ed.) (2018), Handbook of Item Response Theory, Chapman and Hall/CRC, New York, NY.

D. Weiss and A. Sahin (2024), Computerized Adaptive Testing: From Concept to Implementation, Guilford Publications.

Using Conjoint to Assess the Value of Brand Associations

Marco Vriens
Felix Eggers
Kwantumlabs.ai

Introduction

To measure brands via choice-based conjoint analysis or binary choice experiments to understand their impact on consumers has always been somewhat of a challenge (e.g., Guyon and Petiot, 2014). There are at least two approaches in use. One, we can include brand names as an attribute in conjoint analysis (e.g., Dean, 2004; Ferjani, Jedidi and Jagpal, 2009; and Pitcher and Chirilov, 2023), and just like other attribute levels, the brands will get a utility, and these utilities can be used to derive the relative importance of brands, and it can be used to estimate a dollar value of each included brand. There are multiple ways to calculate this value (Befurt, Eggers, Hauser, 2025) but a popular approach is similar to how the willingness to pay for features is determined, e.g., dividing the brand utility by the utility for price, and expressing the resulting brand value relative to a fictitious brand with no value. This dollar value is then the brand's equity. Two, conjoint has also been used to analyze the impact of brand perceptions on the brand's equity. Perceptions cannot be included directly in the conjoint as discrete attributes. For example, Vriens and Frazier (2003), develop an HB conjoint model where the brand utility is modeled as a function of the brand perceptions. The analysis can be done in one integrated step but the perception data are collected outside the conjoint hence not part of the experimental design. This approach works well but has several disadvantages: 1) brand perceptions have a low to non-existent correlation with market share (see Vriens, Chen and Schomaker, 2019), 2) they can be cumbersome to collect as sometimes these brand association lists can become quite long and hence tedious for respondents, 3) brand rating questions may suffer from Halo effects or be vulnerable to response style effects such as straight-lining (e.g., Sonnier and Ainslie, 2011), and, lastly, typically no insight is gained in negative attributes. An alternative approach to get insight into brand associations is through the use of open-ended questions (Vriens, Chen and Schomaker, 2019).

In this paper we aim to:

1. Show the use of open-ended questions to derive the $ value of brand associations.
2. Compare the $ values of open-ended associations with associations based on brand perception ratings.

A New Methodology

Research has shown that brand associations derived from open-ended questions are strongly correlated with market share (more precisely: a brand equity metric derived from these associations is strongly correlated with market share). Open-ended questions don't suffer from Halo and response style effects and will reveal both positive and negative associations (Vriens, Chen and Schomaker, 2019).

The open-ended question used for this can look something like this: "What comes to mind when you think of brand X." This question can be asked in a variety of different ways, but the essence is always the same: simply knowing what respondents think of when they think about a given brand. The open-ended responses are then manually analyzed to develop a brand association dictionary. This results in a set of distinct associations for each brand. Increasingly, this coding can be done using AI. In a recent study we found associations for the Toyota brand such as Practical, Reliability, and Wise Choice. This initial list can be quite long (often more than 100 distinct associations). Then subsequently the data set is scored on the identified distinct associations using advanced AI/text analytics models such as the BERT model (Devlin et al., 2018). This scoring allows us to identify the distinctiveness and strength (relative frequency) of the associations.

Integrating (conjoint) choice data and open-ended response can be done in several ways, e.g., using the graded membership model as used by Liu, Kurz and Allenby (2023). A more direct approach is possible. We replace in the analysis design matrix each brand with only the specific associations mentioned by the respondent in a binary format, i.e., associated (1) or not (0). In a typical brand density study, we find a lot of associations with a given brand. Most will be mentioned by only a few people. So, here we only use the most frequently mentioned associations. We can also add the *number* of distinct associations mentioned by the respondent (as a holistic brand component). We know from brand research that more associations mean more equity. Analytically, this is similar to what we do in the holistic conjoint method (Vriens and Eggers, 2025) where we showed that the holistic feature dimension has a very strong impact on consumer choices. That way, we can add the associations as predictors in a choice model and derive a utility for them. An alternative procedure is to use an HB estimation procedure to obtain individual-level brand utilities and then regress the utilities on the open-ended associations.

In our empirical studies we compare this proposed approach to a conjoint approach that integrates traditional ratings of brand perceptions.

Empirical Studies

We use two studies to evaluate and illustrate the usefulness of our approach.

Study 1. Using Open-Ended Questions to Capture Brand Associations

In study 1, we collected data on several brands. In addition to the open-ended brand density question, respondents were given a single consumer surplus binary choice question, i.e., whether to keep access to the products and services of a certain brand or to give up these goods for 1 month in exchange for monetary compensation. The monetary amounts were experimentally varied between $1 and $1,000 (see Eggers et al., 2024). This question relates to a more complex conjoint method as it only varies price levels for any given brand. The corresponding choice model then models the willingness-to-accept monetary compensation to give up the brand (which is conceptually different to willingness-to-pay measures).

When we estimate the basic consumer surplus model, we find a median consumer surplus of $ 711.50 for Apple and $ 98 for Microsoft, i.e., consumers require a higher compensation to give up Apple products and services than those by Microsoft. When we now add the top associations for each brand from the open-ended questions (note that these associations differ by brand), we get the results shown in Table 1 and 2. There are two columns: One column shows the consumer surplus for those who did not associate the brand with a given association, the second column shows those who do.

Table 1: The Value of Brand Associations for Apple (in $):

	Those who associate	Those who don't associate
Sleek	$ 3215	$ 595
Convenient	$ 2049	$ 591
Innovative	$ 2957	$ 618
Overpriced	$ 84	$ 759

As we can see in Table 1, the top associations add quite a lot of surplus value to the brand, and negative associations reduce the consumer surplus. In Table 2, we see again that associations for Microsoft add $ value to the brand, but not quite as much as for Apple.

Table 2: The Value of Brand Associations for Microsoft (in $):

	Those who associate	Those who don't associate
Job	281	91
Productivity	303	92
Greed	73	100

For each brand we also found a positive effect of the number of associations such that those who have more associations with a brand exhibit higher surplus value. We will show detailed analysis on this in study 2.

Study 2. Traditional Brand Ratings Versus Open-Ended Associations

In study 2 we collected data using the choice-based conjoint approach. The category was premium online job search websites. The conjoint had four attributes: brand, price, salary assessments and skills assessment. Brand had four levels: LinkedIn, ZipRecruiter, Indeed and a fake brand Laboraeus. Price had four levels: $10 per month, $20 per month, $30 per month, and $40 per month. The remaining two attributes were binary: either present or not. Respondents also answered an open-ended brand density question for each of the four brands that were included in the brand attribute and responded to several pre-defined brand perception attributes. This enables us to compare our new approach using open-ended brand associations with the traditional approach of rating the brands on pre-defined scales and integrating these measures in the estimation.

Survey

The structure of the survey was:

1. Open-ended brand association question, one for each of 4 brands,
2. A section where brands are being rated on several associations using a 5-point rating scale, and
3. A choice-based conjoint with three real brands + one fake brand, price + two additional features.

Model Approach

Our modeling approach was straightforward:

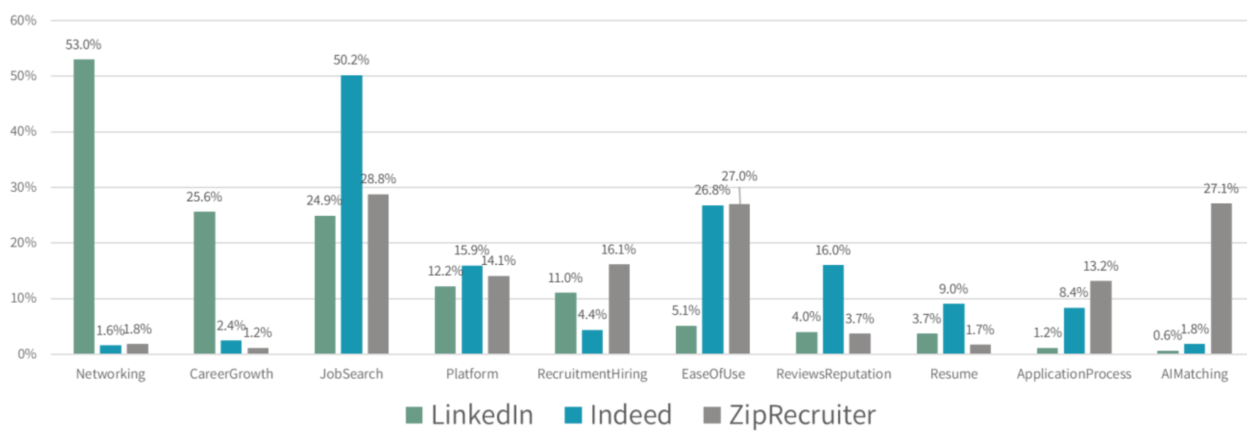
1. Analyze the open-ended responses using automatic text analysis to extract distinct associations for each brand and each participant.
2. Integrate the association matrix with the experimental design matrix to include binary associations, i.e., associated (1) or not (0).
3. Run a choice model in which the distinct binary associations serve as independent variables. Alternatively, for comparison we do the same using the pre-defined brand perceptions measured on the 5-point rating scale.

RESULTS

Data was collected via Prolific. Our sample was net, n=955, 48% were male, 52% were female. Average age was 38.9 years (standard deviation = 12.4).

The analysis of the open-ended brand association questions generated over 80 distinct associations for each brand. Most of these associations are mentioned by five or less respondents and hence cannot be used for further analysis. The chart below shows the most frequently mentioned associations across brands.

Figure 1: Top Extracted Open-Ended Brand Associations across Brands



The first step in our modeling is running the conjoint model to obtain the utilities. See Table 3 for the average utility values of the attributes (for brand, the fake brand served as the reference category with an assumed utility of 0).

Table 3: Conjoint Estimates

	Estimate	Sign.
Brand LinkedIn	0.40	0.00
Brand Indeed	0.15	0.00
Brand ZipRecruiter	0	N.S.
Skills Assessment	1.00	0.00
Salary Insights	1.10	0.00
Price (linear)	-0.02	0.00
None option	1.10	0.00

From this we can calculate the $ values of brand, relative to the fake brand, which has no brand value. In this case: The LinkedIn brand is worth $19.45 per month, and Indeed, $7.23 per month more than the fake brand. Note: these $ values, just like utility values cannot be interpreted in absolute terms. Hence, we included the fake brand which we know doesn't have any brand value.

In Tables 4 and 5 we show the results of the choice model that integrates the most frequently mentioned open-ended associations (Table 4) and pre-defined brand ratings (Table 5).

Table 4: The Impact of Open-Ended Associations on Brand Choice

Open ended associations	Estimate	Sign.
Job Search	0.12	0.01
Networking	0.18	0.01
Ease of Use	0.28	0.00
Career Growth	0.43	0.00
Resume	0	N.S.
Recruitment Hiring	0.14	0.04
AI Matching	0.16	0.05
Application Process	0	N.S.
Platform	0.15	0.02
Reviews Reputation	0	N.S.

Table 5: The Impact of Pre-Defined Perceptions on Brand Choice

Ratings	Estimate	Sign.
Trustworthy	0.10	0.00
User friendly	0	N.S.
Provides relevant recommendations	0.12	0.00
Makes job searching less stressful	0.12	0.00
I enjoy using	0.15	0.00

The estimates in Table 4 show that associating a brand with career growth has the largest impact on choice, followed by ease of use, networking opportunities, and AI matching functions. Of the association ratings in Table 5 we see only minor differences between the dimensions, with enjoyment to use having the largest marginal effect.

Next, we convert these coefficients from Tables 4 and 5 to calculate the $ equivalent. The results of this conversion for both open-end associations and pre-defined perceptions are shown in Table 6.

Table 6: The $ Value of the Top 4 Most Valuable Brand Associations

Open associations	$ value	Pre-defined perceptions	$ value
Career growth	$ 20.5	Enjoy using	$ 8
Ease of use	$ 13	Less stressful	$ 6
Networking	$ 8.5	Relevant recommendations	$ 5.5
AI Matching	$ 7.5	Trustworthy	$ 5

The $ values in Table 6 imply that consumers are willing to pay $20.50 more per month for those brands that they associate with career growth, compared to brands that do not have this association. The remaining associations can be interpreted similarly. These interpretations appear more intuitive than the $ values for the pre-defined perceptions, which imply that one scale-point increase in enjoyment, for example, leads to a higher willingness-to-pay of $8.

Mental Availability

According to recent brand research, mental availability is a key driver in brand equity. Mental availability is a function of distribution and salience (e.g., bright colors) but also how strongly the brand is represented in consumers' minds. The latter can be measured by looking at how many unaided associations respondents mention, or in the case of pre-defined perceptions, how many perceptions receive top 2 box scores (the results of the pre-defined perceptions are not shown here but are similar to the open-ended associations and confirm the mental availability theory). In Table 7 we show the $ value of the number of associations (regardless of what these associations precisely are).

Table 7: The $ Value of Mental Availability

Number of open-ended associations	WTP
0	$ 0
1	$5.00
2	$14.30
3	$24.87
4	$28.42
5	$36.31

Conclusion

In this paper we have shown how we can calculate the $ value of associations when these are elicited with an unaided open-ended question. The results show that the $ value brand associations can vary dramatically. We did a similar analysis with an aided list of pre-defined perceptions. It is interesting to note that unaided brand associations are very different than the typical pre-defined list of perceptions developed by brand managers. Both types of results are useful. The unaided open-ended associations are useful because they show how consumers really think of your brand and how valuable the associations are that the brand knowingly or unknowingly has built in consumers' minds. The aided perceptions are useful because they allow brand managers to test their hypotheses as to what matters to consumers. Both types of results can be used to inform what types of associations the brand wants to build with their messaging.

We also showed that mental availability matters a lot. In addition to building the most valuable associations, brands should also try to build as many associations as they can. The results of this research confirm and extend the results by Vriens, Chen and Schomaker (2019).

Marco Vriens

Felix Eggers

REFERENCES

1. Befurt, D., Eggers, F., & Hauser J.R. (2025). Supply Side Considerations When Using Conjoint Analysis in Litigation. Handbook of Marketing Analytics: Methods and Applications in Marketing, Public Policy, and Litigation, (Edward Elgar), Natalie Mizik and Dominique Hanssens, Eds.
2. Dean, D.H. (2004). Evaluating potential brand associations through conjoint analysis and market simulations. Journal of Product & Brand Management, 13, 7, 506–513.
3. Devlin, J., Chang, M-W., Lee, K. & Toutsnova, K. (2018). BERT: Pre-training of Deep Bi-directional Transformers for Language Understanding. arXiv, preprint arXiv:1810.04805.
4. Eggers, F., Vriens, M., Verhulst, R., Talwar, J. & Collis, A. (2024) Why You Should Be Tracking Customer Surplus Value. Harvard Business Review.
5. Ferjani, M., Jedidi, K. & Jagpal, S. (2009). A conjoint approach for consumer and firm-level valuation. Journal of Market Research, XLVI, 846–862.
6. Gyon, H. & Petiot, J-F. (2014). New conjoint approaches to scaling brand equity and optimizing share of preference prediction. International Journal of Market Research, 57, 5, 701–725.
7. Liu, Y-C. M., Kurz, P., and Allenby, G. (2023). Archetypal analysis and product design. In Sawtooth Software Conference Proceedings, Orlando, FL., 255–267.
8. Pitcher, J. & Chirilov, A. (2023). Harnessing the power of conjoint analysis to track and build brand premium. In Proceedings of the Insights and Analytics Conference, Barcelona, 55–64.
9. Sonnier, G., & Ainslie, A. (2011), Estimating the value of brand-image associations: The role of general and specific brand image. Journal of Marketing Research, 48, 3, 518–531.
10. Vriens, M. & Frazier, C. (2003). The hard impact of the soft touch: How to use brand positioning attributes in conjoint analysis. Market Research Magazine, pp. 22–27.
11. Vriens, M., Chen, S. and Schomaker, J. (2019). The evaluation of a brand association density metric. Journal of Product and Brand Management, 28, 1, 104–116.

12. Vriens, M., Mills, D. & Eggers, F. (2024). Holistic conjoint. In: Proceedings of the Analytics & Insights Conference, San Antonio.
13. Vriens, M. & Eggers, F. (2025). Holistic Conjoint. Customer Needs and Solutions, vol. 12, no. 8.

AI-ASSISTED SEGMENTATION: BETTER, FASTER, AND MORE ACTIONABLE SEGMENTATION

JACKIE GUTHART
CURTIS FRAZIER
MARCOS NUÑEZ
YAMIL BONGIOANNI
MARISSA SIMONS
RADIUS GLOBAL MARKET RESEARCH

ABSTRACT

Segmentation is a cornerstone of market research, driving strategy across acquisition, retention, and targeted activation. Yet the traditional segmentation workflow is often inefficient and prone to cognitive overload. This paper presents an AI-assisted hybrid workflow developed at Radius Global Market Research that dramatically accelerates solution evaluation, enhances interpretability, and supports activation readiness. We detail our proprietary segment scoring methodology, demonstrate AI-powered segment description and comparison capabilities, and provide a case study from a health and wellness segmentation project. Our results show a >10x efficiency improvement with enhanced quality and actionability.

INTRODUCTION

Segmentation analyses help businesses deeply understand customer needs, attitudes, behaviors, and motivations. The resulting insights drive acquisition, optimize media spend, inform product development, and deepen loyalty.

However, traditional segmentation workflows face key challenges:

- Dozens of potential solutions must be manually evaluated.
- Human judgment is taxed by inconsistent and overwhelming data volumes.
- Segment profiles require time-intensive crafting.
- Solution comparisons can lack transparency and rigor.

The outcomes of these challenges are two-fold. First, the inefficiency of the process means that the analyst spends significant time on solutions that are unlikely to be evaluated as high quality. Second, because of this inefficiency, the analyst may not be able to devote enough time to fully uncovering the patterns within the higher quality solutions. By using a mix of programmatic rules and AI assistance, we propose a new process to alleviate these challenges.

Our objective was to build a hybrid workflow where AI enhances, but it does not replace human expertise. We pair algorithmic solution scoring with AI-generated segment descriptions and comparisons, enabling our analysts to focus on strategic interpretation.

WORKFLOW OVERVIEW

Our AI-assisted segmentation workflow comprises three key components:

1. **Segmentation Development Assistant**
   - Proprietary scoring algorithm prioritizes solutions for deeper analysis.
2. **AI Segment Evaluation and Naming Tool**
   - Generates segment names, descriptions, key attributes, and demographic insights.
3. **AI Segment Comparison Assistant**
   - Facilitates transparent, actionable comparisons of competing solutions.

Each component is detailed below.

SEGMENTATION DEVELOPMENT ASSISTANT

The purpose of this tool is to efficiently evaluate dozens of segmentation solutions using a transparent, repeatable scoring algorithm that is customized to what we know about how the client is likely to evaluate the solutions. By customizing the algorithm based on *a priori* knowledge about our client's needs and preferences, we avoid the pitfall of a one-size-fits all approach. The end goal is a one-number "quality" score that allows the analyst to prioritize those solutions that are most likely to meet our client's needs and spend less time evaluating solutions that are less likely to do so.

Segment Scoring Methodology

The Segment Score combines three components:

1. Segment Distribution Score
2. Average Variance Score
3. Differentiation Score

1. Segment Distribution Score

Our goal by including the segment size distribution into the algorithm is to identify and avoid segmentation solutions with highly skewed segment sizes and/or segments that are too small to be useful to the client. Note that our goal is not necessarily to achieve perfectly evenly-sized segments. Rather, we want a metric that measures and alerts us to distributions that are so uneven that the solution is unlikely to be high quality.

The distribution score can be operationalized in two ways. First, it can be given a weight to be incorporated into the final "quality" score. An alternative is to identify a cut-off point at which a distribution is too skewed to be considered worth further investigation. In this case, we can mark the solution as disqualified.

Formula (Weight Component):

Let $p(s_i)$ = proportion in segment i, N = number of segments.

$$p_{ideal} = \frac{1}{N}$$

$$\text{Distribution Score} = 1 - \sum_{i=1}^{N} |p(s_i) - p_{ideal}|$$

Disqualifier Rules:

A solution is disqualified if either:

- Minimum segment size $n_i \geq 100$ respondents is violated.
- "Proper distribution" fails: Where Y = 2.5

$$\max\left(\frac{p_{\max}}{p_{\min}}\right) \leq Y$$

Please note that these specific cut-off values (n>=100 and Y=2.5) can be modified based on total sample size and number of segments being retained in the analysis. Additionally, we have found with models calling for a large number of segments, that the Disqualifer Rules actually kick-in too often, thereby necessitating making it optional.

2. Average Variance Score

The goal of including average variance is to maximize variance across priority variables. Again, it is vital to note that the analyst is identifying and weighting the "priority" variables based on feedback from the client about goals, expectations and intended uses of the segmentation research.

Note that in order to calculate this variance score, any categorical variables must be converted into a series of dichotomies for the profiles.

Steps:

1. Select priority variables and assign weights w_j.
2. For each variable j, compute variance of segment means Var_j.
3. Compute weighted average variance:

$$\text{Weighted Average Variance} = \frac{\sum_j w_j \cdot \text{Var}_j}{\sum_j w_j}$$

4. Normalize:

$$\text{Normalized Variance Score} = \frac{\text{Weighted Average Variance}}{\text{Maximum Expected Variance}}$$

3. Differentiation Score

Finally, we want to include the Differentiation Score to reward segments that meaningfully "win" on key variables. This is done for a couple of reasons. First, we have found that clients understand and resonate more with segments in which each segment "stands out" on multiple attributes—as opposed to a segment being average across everything. Second, when looking forward to developing a typing tool, we have found that "winning" or "standing out" on key variables significantly increases the prediction rate of our algorithms.

Steps:

1. Calculate each variable's Index-Differentiator score.
   - Use custom formulas for color-coding index scores that both take into account magnitude of index and absolute means. Really high index scores with low total population means may receive lighter green color coding vs. really high index scores and moderate-higher sample means which will receive the darkest green.
   - Calculate an Index-Differentiator score to emphasize over-indexing, meaning we want to draw more attention to where the segment is performing high on an attribute.

		Index-Differentiator Score
	neutral/at par	0
	a little above average	1
	a little below average	0.5
	somewhat above average	2
	somewhat below average	1
	moderately above average	4
	moderately below average	2
	very above average	8
	very below average	4

Over-index (green) = higher weight

Purple (high under-index): lower weight

2. For each priority bucket, compute:

$$\text{Winning }\%_k = \frac{\text{Number of priority variables where segment wins}}{|B_k|}$$

3. Next compute the Initial differentiation score: Let $d_1, d_2, \ldots, d_n$ the differentiation scores for each of the n segments (e.g., the percentage of wins on priority variables). Then:

$$\text{Initial Differentiation Score} = 1 - \frac{\sigma(d_1, d_2, \ldots, d_n)}{\sigma_{\max}(n)}$$

   Where:
   - $\sigma(d_1, d_2, \ldots, d_n)$ is the standard deviation of the differentiation percentages across the n segments.
   - $\sigma_{\max}(n)$ is the maximum possible standard deviation for n segments, used for normalization.

4. Let S_{init} = Initial Differentiation Score and Winning $\%_k$ is defined above as percent of priority variables for which any segment is a "winner" within bucket k:

$$\text{Final Differentiation Score} = \frac{1}{2}\left(S_{\text{init}} + \text{Winning}\,\%_k\right)$$

Final Segment Score

Our final segment "quality" score is simply a weighted average of the three components of the index. The specific formula is:

- w_{dist} , w_{var} = global weights for Distribution and Variance Scores
- $w_{diff,k}$= weight assigned to priority bucket k (with k=1,2,…,)
- D_k = Differentiation Score for bucket k

$$\text{Final Segment Score} = w_{\text{dist}} \cdot \text{Distribution Score} + w_{\text{var}} \cdot \text{Variance Score} + \sum_{k=1}^{n} w_{\text{diff},k} \cdot D_k$$

The weights assigned to each of the three components are scaled such that they will sum to 1.

AI Segment Evaluation and Naming Tool

The second leg of our approach, the AI Segment Evaluation and Naming Tool, automates the generation of rich segment profiles and names. While not a replacement for an analyst, the tool gives a snapshot of the interpretation of each segment and provides several suggested segment names. This snapshot allows the analyst to understand the general outline of the segments. This, in turn, has two primary benefits. First, it means that the analyst can spend more time with those solutions that align with what we know about our client's needs and those solutions that meet face validity requirements. Second, the paragraph-based descriptions (when combined with human interpretation) are easier for most clients to understand than spreadsheets full of values.

How is this implemented? The most important step is to convert our numbers-based profiles into a format that the AI can read. (NOTE: the pace of AI development is rapid. At the time of writing, AI has difficulty with tables of raw values and indices. This may change by the time of publication.)

To create a readable format, we convert our index score color-based conditional formatting into text-based index tables (ranging from Extremely Positive to Extremely Negative). AI has difficulty interpreting colors so we convert the color-coding to text.

Next, our software development team has developed an API into our online segmentation tool that calls a predefined ChatGPT prompt. That prompt requests that the GPT:

Acts as an expert marketing analyst specializing in segmentation and pattern recognition

1. *Generate 4 potential names for each segment.*
2. *Write ≤600-word description.*
3. *Cite row numbers and key attributes.*

This last request is vital as it allows the analyst to verify from where in the segment profiles the AI is pulling its conclusions.

Example Output:

Segment Naming Options: Organic Wellness Enthusiasts, Holistic Health Advocates, Ingredient-Conscious Millennials, Experimental Supplement Pioneers

*Summary: This segment is **highly knowledgeable, value-driven, and experimental** in their health and wellness approach. They seek **organic, non-GMO, and whole-food-based supplements**, and they favor brands that **prioritize transparency and ethical sourcing**. They **actively explore new wellness trends** and focus on **holistic health**, incorporating mental and emotional well-being into their approach. Their purchasing decisions are **not influenced by traditional medical endorsements or price**, but rather by **alignment with personal values and ingredient transparency**.*

Demographics: Millennials, Females, $75k income, suburban

Comparison of Human Description vs. AI Descriptions:

To test the quality of the AI generated summaries, we had our segmentation analysts manually flag areas of emphasis based on segmentation means and index tables and compared the overlap to the call-outs from the AI description assistant. There is a 94.5% overlap.

- AI Strengths: Identifying and flagging particularly low scores; identified some scores overlooked by human evaluation
- AI Weaknesses: Consistency across columns; emphasis on relatively high index values with low total sample means

AI Segment Comparison and Deep Dive Assistant

The final piece of our proposed process is the AI Segment Comparison and Deep Dive Assistant. The AI assistant reads in AI-descriptions, segment averages, and index files, and evaluates key similarities and differences between segments. This allows us to understand how different solutions compare to one another.

Prompt Example:

You are a highly skilled data analyst specializing in market segmentation. Your task is to compare multiple segmentation solutions and recommend the most actionable and insightful approach for the client's needs.

You will analyze:

- *Segment descriptions (attitudes, behaviors, demographics)*
- *Mean scores and index tables (priority variables and differentiation)*
- *Client priority lists (to weigh variables appropriately)*

Process:

1. *Clarify solution names and client context.*
2. *Prioritize higher-importance variables; avoid over-weighting low-importance factors.*
3. *Compare solutions on:*
   - *Differentiation and uniqueness of segments*
   - *Alignment with client targeting needs*
   - *Ease of activation and business relevance*
4. *Create:*
   1. *Closest-match segment comparisons across solutions*
   2. *Quality assessment: Strengths and weaknesses of each solution*
   3. *An activation readiness score for each*
   4. *Segments the client should target for being high-value/opportunity*
5. *Ask clarifying questions as needed (e.g., priority segments, variables, format preferences).*

Final Output:

A clear, data-driven recommendation of the most actionable segmentation solution for the client's goals, with supporting rationale and trade-offs.

There are several pieces of output that the AI is asked to generate.

1. **Closest Matching Segment:** This is an estimate of how each segment in Solution X matches with each segment in Solution Y.

Solution 1 Segment	Solution 4—Closest Match	Key Similarities	Key Differences
Segment 1: Holistic Health	Segment 5: Organic Wellness	Both segments emphasize transparency, holistic health, and ingredient purity.	Solution 1 segment has a stronger focus on scientific credibility, while Solution 4 leans towards ethical sourcing.
Segment 2: Flavor-Driven Supplement	Segment 1: Practical Health	Preference for taste, convenience, and ease of integrating supplements into daily routines.	Solution 1 emphasizes variety in flavors, whereas Solution 4 is more about convenient, no-fuss consumption.
Segment 3: Trusted Supplement Saviors	Segment 2: Trusted Brand	Strong brand trust, reliance on medical recommendations, and preference for credibility.	Solution 1 consumers place higher trust in professional recommendations, whereas Solution 4 allows for more flexibility.
Segment 4: Value-Conscious Supplement	Segment 4: Value-Conscious Supplement	Highly price-sensitive consumers, prioritizing affordability over trends.	Solution 1 buyers are more research-driven, while Solution 4 buyers prioritize straightforward choices.
Segment 5: Organic Wellness	Segment 5: Organic Wellness	Both segments have a strong interest in organic, natural, and holistic wellness products.	Solution 1 consumers are more proactive and knowledge-seeking, whereas Solution 4 consumers are more influenced by emerging trends.

In the past we've done simple crosstabs of segment membership to see movement of respondents from one solution to another. One caution on using AI—the AI will always try to provide a "best" answer whether or not there is a clear matching segment. For example, combining the table above with the crosstabulation below, we can see that the AI correctly matched most of the segments between the two solutions, but mismatched two segments that didn't really have a close analogue between the two solutions.

		Solution 1						
		Seg 1	Seg 2	Seg 3	Seg 4	Seg 5	Seg 6	Total
Solution 4	Seg 1	5%	60%	3%	10%	8%	38%	748
	[illegible]	24%	5%	60%	19%	10%	13%	991
	[illegible]	30%	7%	12%	9%	28%	15%	787
	Seg 4	20%	15%	17%	60%	5%	20%	1158
	Seg 5	21%	13%	8%	1%	49%	13%	816
	Base: Total Respondents	779	471	678	1068	914	590	4500

Solution 1 Segment	Solution 4 – Closest Match
Segment 1: Holistic Health	Segment 5: Organic Wellness
Segment 2: Flavor-Driven Supplement	Segment 1: Practical Health
Segment 3: Trusted Supplement Saviors	Segment 2: Trusted Brand
Segment 4: Value-Conscious Supplement	Segment 4: Value-Conscious Supplement
Segment 5: Organic Wellness	Segment 5: Organic Wellness
Segment 6: Value-Conscious Wellness	Segment 4: Value-Conscious Supplement

2. **Quality Assessment:** The AI can provide guidance on the strengths and weaknesses of each solution. Example output may look like this:

Solution 1 Strengths

- ✅ **Clear Differentiation**
- ✅ **Strong for Marketing & Sales**
- ✅ **Scientific Emphasis**
- ✅ **Psychographic Storytelling**
- ✅ **Long-Term Strategic Fit**

Solution 4 Strengths

- ✅ **Simpler Segmentation**
- ✅ **Broad Consumer Reach**
- ✅ **More Behavioral-Focused**
- ✅ **Easier Internal Adoption**

Solution 1 Weaknesses

- ❌ **More Complexity**
- ❌ **Internal Adoption Challenges**
- ❌ **Potential Over-Segmentation**

Solution 4 Weaknesses

- ❌ **Less Differentiation**
- ❌ **Weaker Scientific Emphasis**
- ❌ **Less Precise Targeting**
- ❌ **Limited Long-Term Strategic Fit**

3. **Activation Readiness Score:** This is an example of where the AI acted as a "thought partner" to develop an unexpected output. When we had ChatGPT help us perfect the prompt for the comparison assistant it came up with this addition, came up with the metrics, and the weights for this project.

Metric	Weight	Solution 1	Solution 4
Ease of Ad Targeting	25%	9	7
Sales & Distribution Usability	10%	9	7
Psychographic Storytelling Potential	20%	9	7
Simplicity for Internal Teams	15%	7	9
Direct Link to Product Features	20%	8	6
Long-Term Strategic Usability	10%	8	6
Activation Readiness Score		**8.4/10**	**7.0/10**

4. **Estimate Segment Value: Identifying the Best Consumer Segments for Marketing, Growth, and Spend:** While it is common to develop a programmatic assessment of segment value, the AI was able to generate an estimate on its own. Below is an example of how AI operationalized its assessment of segment value:

Top Spending Segments 💰

Holistic Health Millennials—Segment 1

- *High-income consumers ($100K+ HH income)*
- *Frequent supplement users (wellness drinks, protein powders, greens powders)*
- *Premium brand preference and scientific credibility focus*

Organic Wellness Enthusiasts—Segment 5

- *Willing to pay premium for clean, organic ingredients*
- *Heavy supplement users (collagen, probiotics, herbal, wellness drinks)*
- *Prioritize ingredient transparency and sustainability*

Follow-up questions to the Segment Comparer Assistant: While many of the outputs listed above could be generated through hours of manual evaluation of segmentation solutions, the AI can also be used to provide information unavailable in our survey data.

1. **Omitted Survey Topics:** In almost every survey, there are metrics that, either because of survey length restrictions or simple oversight, are omitted from the research. We have found that the AI can be used to fill in some of those gaps. For example, in our research, we did not ask about media usage. If we want an estimate of how each segment would be *likely* to use different media platforms, the AI can combined demographic profiles from the survey with publicly available usage data:

Prompt: "What platforms are most effective for marketing/advertising for the top spending segments?"

Example Output:

Segment 5: Organic Wellness Enthusiasts:

Instagram and Pinterest *→ Aesthetic, natural ingredient-focused storytelling.*
TikTok (Holistic health trends and influencer partnerships) *→ Viral health trends influence their choices.*
YouTube (Sustainability and ethical brand storytelling) *→ Deep dives into sourcing, organic farming, and wellness.*
Sponsorships on Natural Living Blogs and Conscious Consumer Platforms *→ They actively seek ethical brands.*
Marketing: Focus on **organic, non-GMO, ethical sourcing, and sustainability**

2. **Using AI as a creative thought partner:** One of the most significant challenges in segmentation research is to bring the segments alive for clients. One creative idea is to have the AI evaluate the segment profiles and assign a well-known celebrity to each segment.

Prompt: "Which famous people might fall into each segment?"

Example Output:

Segment 5: Organic Wellness Enthusiasts:

Jessica Alba: *Founder of The Honest Company, a strong advocate for clean, organic products.*
Zac Efron: *Has spoken about his switch to holistic, clean eating and natural supplements.*
Alicia Silverstone: *Vegan and organic advocate who strongly promotes ingredient transparency.*

Conclusion

AI accelerates and sharpens segmentation but works best with human review and guidance. This approach delivers better outcomes faster and with more confidence.

Traditional, Human-Only Approach:

Strengths

✓ Stable evaluation criteria across projects

✓ Human judgment understands nuance and scope of full research project

Weaknesses

❌ Time consuming and error prone

❌ Limited by cognitive overload

AI-Assisted Approach Coupled with Segment Scoring Algorithm:

Strengths

✓ Over 10x as fast

✓ Ability to analyze hundreds of variables and several solutions

✓ Provides creative/non-obvious insights

Weaknesses

❌ Sometimes overemphasizes less meaningful patterns

❌ Can hallucinate: requires human oversight

The Best Approach is a Hybrid Model:

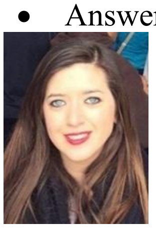
- AI accelerates the segmentation process
- Humans validate, interpret, and refine AI results
- AI + Human teams create faster, more accurate, and actionable segmentations

In Development: Additional AI Segmentation Features

Where is this going next? One of the key requirements to make this hybrid approach function is the quality of the information being provided to the AI. So, wherever possible, we would like to provide additional context (qualitative interviews, other findings from studies, summaries of client discussions) to AI to strengthen the segment description/naming tool/comparison tool.

In addition, we believe that for socialization of the results to our clients, we can include an AI chatbot used for analysis and during client sessions in which the user can ask questions about the segments or competing solutions. These questions may include asking the AI to:

- Provide thought-starters on marketing campaigns to reach target segments
- Guide priorities for future research
- Answer follow-up questions



Jackie Guthart

Curtis Frazier

Marcos Nuñez

Yamil Bongioanni

Marissa Simons

SEGMENTATION 2.0:
REDEFINING SEGMENTATION FOR MODERN MARKETING

CATHERINE GIBSON
JESSICA WOJTUNIK
ALEXANDRA CHIRILOV
AGNIESZKA FRONCZYK
RACHEL THOMPSON
IRINA NAZAROVA
NIQ

ABSTRACT

Modern Marketing theory has evolved and is prompting a critical reassessment of foundational strategic marketing frameworks, particularly the Segmentation, Targeting, and Positioning (STP) model. Lately, researchers and influential scholars have started challenging the primacy of differentiation and targeting, advocating instead for broader market penetration and the cultivation of mental and physical availability. These perspectives raise a fundamental question: Does segmentation still hold relevance in contemporary marketing practice?

The authors contend that segmentation remains indispensable—but must be redefined conceptually to feed into the new marketing paradigm. We propose to expand the use case of segmentation by integrating traditional segmentation with additional analytics to support both short-term sales activation and long-term brand building. Feeding into dual growth strategies we begin with the identification of distinct consumer segments and extend the insights to feed into broader audience engagement through similarity analysis of the segments.

Our approach comprises three analytical extensions to traditional segmentation: (1) *upscaling*—leveraging pairwise similarities to expand from core targets to adjacent segments; (2) *broad reach targeting*—aggregating segments into thematic platforms to enable scalable communication; and (3) *brand building*—identifying universally resonant themes to support mass-market messaging. It reconciles the dichotomy between targeting and mass marketing, offering an empirically grounded and actionable model for sustainable brand growth.

MOTIVATION

This exploration is motivated by a growing theoretical and practical tension within the field of marketing. On one side stands the traditional Segmentation, Targeting, and Position (STP) paradigm, as articulated by Philip Kotler, which emphasizes the identification of distinct consumer segments, targeted communication, and differentiated positioning. On the other, empirical research by Byron Sharp and the Ehrenberg-Bass Institute has demonstrated that brand growth is predominantly driven by market penetration—reaching all category buyers rather than just narrowly defined targets (Sharp et al., 2024). This divergence has catalyzed a polarizing debate: Should marketers prioritize targeted strategies or embrace mass marketing?

As practitioners and developers of segmentation solutions, we found ourselves at the intersection of this debate. We posit that segmentation is not obsolete but must evolve to meet the demands of a dual-strategy marketing environment. Traditional segmentation remains effective for short-term activation by identifying differences and contextual drivers. However, it's insufficient to inform long-term brand building, which requires uncovering commonalities and scaling insights across broader audiences.

This paper introduces a redefined segmentation framework that addresses this gap. By integrating traditional segmentation with additional analytical techniques, we aim to provide a comprehensive model that supports both immediate commercial objectives and enduring brand equity. Our objective is to inspire a new perspective on segmentation that is empirically grounded, analytically enriched, and strategically aligned with the realities of modern marketing.

Evolution of Marketing Theory

The theoretical foundations of marketing have undergone significant transformation over the past decades. From the structured logic of segmentation and differentiation to the empirical generalizations of brand growth, the field has seen a shift from intuition-driven frameworks to evidence-based models. This section examines three pivotal perspectives that have shaped this evolution: Kotler's STP model, Sharp's market-based asset theory, and the empirical work of Binet and Field.

1. Philip Kotler and the STP Paradigm

Philip Kotler's Segmentation, Targeting, and Positioning (STP) model has long served as the cornerstone of strategic marketing. Introduced in the 1960s, the STP framework posits that markets consist of heterogeneous consumers who can be grouped into meaningful segments based on their needs. Marketers are advised to identify these segments, select those whose needs they can serve most effectively, and position their offerings to meet the specific needs of the chosen targets (Lynn, 2012).

Kotler and Keller (2021) describe this process to "deliver high value and satisfaction, which lead to high repeat purchases and ultimately to greater company profitability" (p. 167). The STP model assumes that differentiation and relevance are key to competitive advantage, and that brands can secure loyalty by tailoring their value propositions to narrowly defined audiences.

While the STP model remains widely taught and applied, it has come under increasing scrutiny for its limited explanatory power considering empirical evidence on how brands grow and compete. A prominent critic of this model is Byron Sharp, from Ehrenberg-Bass Institute. In *How Brands Grow* (Sharp, 2010) he argues that brands compete on look-alikes and questions the importance of differentiation.

2. Byron Sharp and How Brands Grow

Byron Sharp, along with colleagues at the Ehrenberg-Bass Institute, has fundamentally challenged the assumptions of the STP model. In *How Brands Grow* (Sharp, 2010) and subsequent empirical studies (Sharp et al., 2024), Sharp argues that brands grow not by deepening loyalty within narrow segments, but by increasing penetration—reaching more category buyers.

Sharp's *How Brands Grow* (Sharp, 2010) is grounded in empirical laws and generalizations such as:

- *The Duplication of Purchase Law*: Competing brands share customers in proportion to their market share.
- *The Double Jeopardy Law:* Larger brands enjoy both more customers and slightly higher loyalty.
- *User Profile Similarity:* Buyers of competing brands tend to look demographically and behaviorally similar.

These findings contradict the STP assumption that brands serve distinct customer bases. Instead, Sharp posits that mental and physical availability—being easy to think of and easy to buy—are the true drivers of brand growth (Sharp et al., 2024).

3. Binet and Field: Balancing the Long and the Short

While Sharp outlines the empirical patterns underlying brand growth, centered on penetration, mental and physical availability, Binet and Field focus on the temporal dynamics of advertising effectiveness. Their framework distinguishes between long-term brand equity building and short-term sales activation, emphasizing differential effects across creative strategy, media channels, and investment horizons.

Binet and Field's research shows that *short-term campaigns* drive immediate sales through targeted, product-focused messaging, while *long-term campaigns* build brand equity through emotional, broad-reach communication.

Although the underlying principles of their work build on earlier research in advertising and media effects, Binet and Field's empirical synthesis has gained broader recognition among practitioners only in recent years. In their seminal work *The Long and the Short of It* (Binet, L., and Field, P., 2013), they argue that optimal marketing effectiveness is achieved through a 60/40 split between long-term brand building and short-term activation.

INDUSTRY TRACTION

Former Marketing Professor and Marketing Consultant, Mark Ritson has been instrumental in translating these findings into strategic guidance for practitioners. Drawing on the empirical work of Binet and Field, as well as Sharp, he advocates for what he terms "Bothism": the simultaneous pursuit of mass brand building and targeted activation. Ritson challenges the dichotomy between mass marketing and targeting, arguing that "the only way to achieve sustainable short- and long-term growth is to balance targeted activation with brand building aimed at the whole market" (Ritson, 2018). His perspective emphasizes the practical necessity of an integrated dual-strategy approach, grounded in the empirical evidence of contemporary marketing science.

Together, these perspectives illustrate the evolution of marketing theory from a focus on differentiation and targeting to a more nuanced understanding of how brands grow. While Kotler's STP model remains foundational, it is increasingly viewed as insufficient in isolation. Sharp's empirical laws and Binet and Field's evidence-based balance of short-term sales activation with long-term brand building all point toward a need for redefining segmentation.

Redefining the Segmentation Framework

The framework presented in this paper builds on this intellectual lineage (See Figure 1). It retains the diagnostic power of traditional segmentation while extending its utility through additional analytics that support broader reach and long-term growth. This duality reinforces the need for a segmentation approach that serves both functions. Traditional segmentation supports short-term targeting, while a redefined approach—proposed in this paper—also enables long-term brand building by identifying themes that resonate across segments.

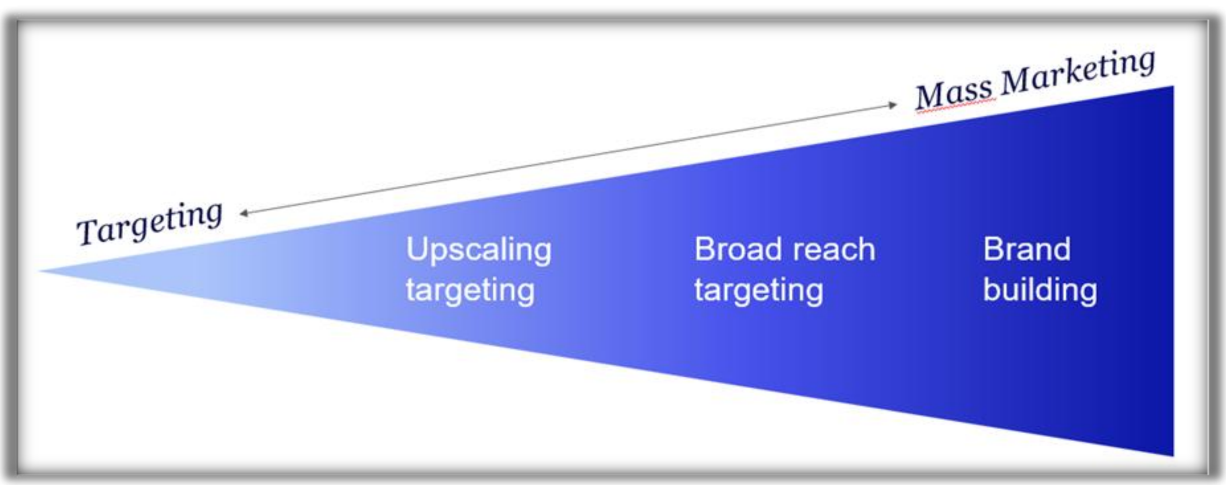
Considering the fundamental shifts in marketing theory, it is evident that segmentation must evolve to remain relevant. Segmentation can serve as a foundation for generating insights that inform both short- and long-term growth strategies by identifying meaningful differences across segments while simultaneously uncovering commonalities to enable scalable marketing efforts.

Traditional segmentation clusters consumers or purchase occasions into distinct groups to facilitate targeted activation. Accounting for consumer heterogeneity enables firms to refine these strategies by focusing on segments that are most responsive to specific appeals. Short-term sales activation operates primarily at the lower stages of the purchase funnel, employing functional or product-centric messages aimed at triggering immediate purchase behavior within targeted segments or consumption contexts.

Consistent with recent shifts in marketing theory, we extend beyond that conventional application of segmentation. By integrating segments through finding their shared characteristics, firms can achieve scalability and support long-term strategic objectives by addressing the broader market at the upper stages of the purchase funnel.

Segmentation serves as a tool to find patterns in the market, that can be used to scale up and down. There are different levels of scale between targeting and mass marketing. A brand's position on this continuum is determined by factors such as firm size, budget constraints, and growth objectives. Utilizing segmentation as a foundation to inform both short-term activation and long-term brand-building efforts ensures that these activities are aligned and are mutually reinforcing within a unified strategic framework.

Figure 1: Expanding the Segmentation Use Case



1. Targeting

Consistent with Kotler's framework, the process begins with traditional segmentation, which remains effective for short-term activation by identifying consumer groups with shared preferences and needs. Analyzing these segments reveals underlying differences and contextual drivers that characterize segment-specific behavior and uncover potential growth opportunities. Detailed segment profiling enables marketers to refine product positioning, develop targeted innovations, and tailor marketing interventions to enhance short-term sales performance.

This step is supported by advanced clustering techniques that identify consumer or occasion segments that meet the client use case. Given the widespread application of this approach, we will not elaborate on its specifics in this context.

2. Upscaling

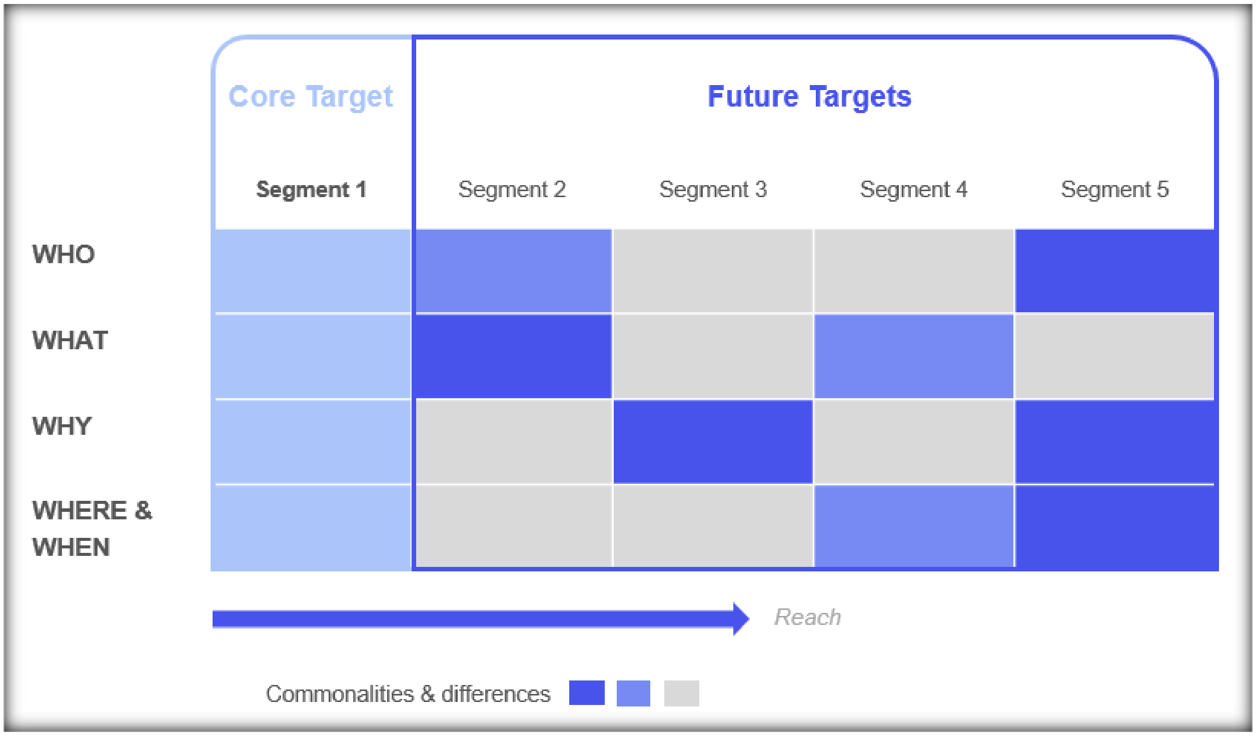
Conceptually, expanding from a core target segment enables sales growth within the primary audience while simultaneously managing the broader market effects of targeted activities. Rather than focusing solely on what differentiates the core target, the objective is to identify the similarities or shared attributes that exist across segments.

Methodologically, upscaling involves identifying *shared themes* that link the core segment with adjacent or future target groups and applying similarity measures (e.g., Euclidean distance) to quantify pairwise similarities between segments. These commonalities can then be leveraged to build brand associations that extend beyond the initial target audience.

Traditional segmentation often assumes that segments are mutually exclusive and neatly delineated, overlooking the fact that meaningful commonalities may exist across different dimensions. Segments may diverge on one attribute while converging on others, reflecting the inherently multidimensional nature of consumer behavior. In the context of occasion-based segmentation, it is therefore recommended to analyze the key situational dimensions—the "big Ws" (see below)—independently, to extract insights from multiple perspectives.

By analyzing pairwise similarities between the core and other segments—a similarity/dissimilarity matrix provides a quick view on how similar the segments are. This is illustrated by the following hypothetical example (see Figure 2). When analyzing the Who dimension, Segment 1 appears very similar to Segment 5 and moderately similar to Segment 2. In contrast, when examining the Why dimension (i.e., consumption motives), Segment 1 shares common reasons with both Segments 3 and 5.

Figure 2: Upscaling Visual



The degree of similarity or difference between segments thus depends on the specific dimension under consideration. Segments may differ in contextual factors while figuring convergence in underlying needs. From this analysis (see Figure 3), three core insights can be derived:

1. Which W offers the greatest potential for expanding reach?
2. Which additional segments may be addressable?
3. Which themes should be prioritized in communication strategies?

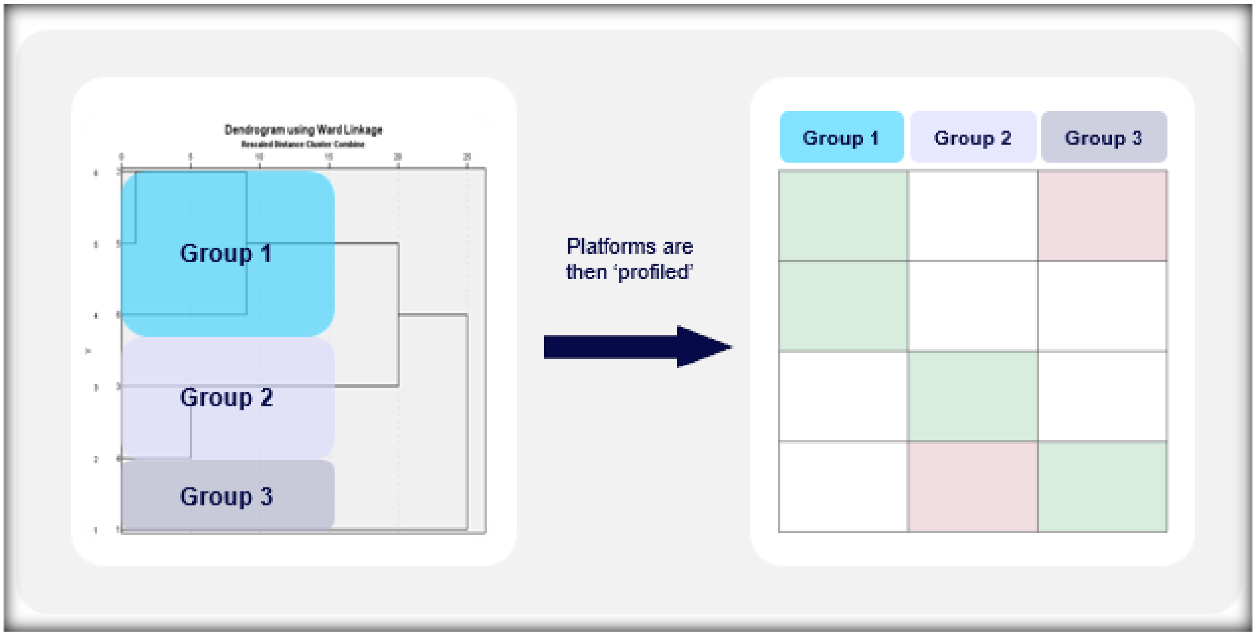
Figure 3: Upscaling from a Core Target Segment

3. Broadening the Reach

In the preceding section, we examined approaches for scaling from a specific segment. However, when the objective is to build broader mental availability and support long-term growth, it becomes necessary to expand the analytical lens beyond individual segments. Identifying broadly resonant themes requires the development of messages that appeal across the entire category, or the creation of thematic platforms that are universally accessible and avoid alienating any subgroup within the market.

Individual segments are aggregated into broader clusters using hierarchical cluster analysis (HCA), based on the same input variables that were employed in the initial segmentation. This second-level clustering groups the segments into larger, more prominent platforms (see Figure 4). Each platform is subsequently profiled to identify the themes that resonate most strongly within the broader groupings. Instead of having several segments to target individually, creating platforms helps to inform positioning strategies that will reach a couple of segments in a way that is meaningful to all of them.

Figure 4: Platform Development Across Segments



4. Long-Term Brand Building

Long-term brand building builds upon the previously outlined approach but extends it by identifying universal themes that resonate broadly across all segments, thereby supporting mass-market messaging. For example, if one's focus was on the non-alcoholic beverage market, messaging such as "*taste I love*" and "*refreshing*" may exemplify the type of universally appealing themes that transcend individual segment boundaries.

In this discovery phase, we evaluate the leading themes from a total market perspective, while simultaneously considering segment-specific performance. It is important to note that when profiling, a high mean across the entire sample doesn't necessarily reflect widespread engagement across all groups; concentration within specific subgroups may inflate aggregate metrics without ensuring broad-based resonance.

To address this, we applied the *MaxiMin rule* strategy introduced by the American philosopher John Rawls, which focuses on maximizing the minimum payoff across alternatives (John Rawls, 1971). In this context, we identify topics that deliver the highest net positive impact while ensuring the broadest positive reception across segments. This approach minimizes the risk of alienating any segment by selecting themes that yield the greatest collective reach while safeguarding against significant negative reactions in any segment.

Summary

This paper introduced an extended framework for segmentation that incorporates three complementary analytical approaches: Upscaling, Broad Reach, and Brand Building. Expanding traditional segmentation, which primarily emphasizes differentiation between groups, this framework adds new layers of analysis that identify not only what separates segments, but also their underlying commonalities. The key insight emerging from this research is that identifying differences is as important as recognizing commonalities. The creation of distinct segments serves as a necessary starting point to systematically uncover shared attributes that inform broader marketing strategies.

By expanding the use case of segmentation in this way, the proposed approach aligns with the evolving demands of modern marketing theory, which increasingly seeks to reconcile the tension between targeted activation and mass-market brand building. Rather than positioning targeting and mass marketing as opposing strategies, this framework enables both to coexist within a unified structure. The ability to derive segment-specific insights for short-term activation, while simultaneously identifying universal themes for long-term brand building, offers a more flexible and scalable approach to market segmentation.

Conceptually, this expanded framework reflects a shift in segmentation from a static classification tool to a dynamic analytical process that feeds into the broader marketing paradigm. It supports the development of positioning platforms that can scale from niche audiences to the broader market without sacrificing relevance or reach. Moreover, by integrating methods such as similarity statistics, hierarchical cluster analysis, and decision criteria inspired by game theory (e.g., MaxiMin rule), this approach enhances the analytical rigor of segmentation practices.

Catherine Gibson

Jessica Wojtunik

Alexandra Chirilov

Agnieszka Fronczyk

Rachel Thompson

Irina Nazarova

References

Binet, L., & Field, P. (2013). The Long and the Short of It. IPA.

Kotler, P., & Keller, K. L. (2021). Marketing Management (16th ed.). Pearson.

Lynn, M., (2012). Segmenting and Targeting Your Market: Strategies and Limitations. The Cornell School of Hotel Administration on hospitality: Cutting edge thinking and practice, pp. 351–369.

Rawls, J. (1971). A Theory of Justice. Belknap Press of Harvard University Press

Ritson, M. (2018). Targeting or mass marketing? The answer is both. Marketing Week.

Sharp, B. (2010). How Brands Grow. Oxford University Press.

Sharp, B., Dawes, J., & Victory, K. (2024). The market-based assets theory of brand competition. Journal of Retailing and Consumer Services, 76, 103566.

MODERNIZING DATA VISUALIZATION PRACTICES FOR MARKET RESEARCH

J. KEATON WILSON
BEN CORTESE
KS&R

INTRODUCTION

Market researchers today face an abundance of data but often rely on outdated visualization habits. Much of the emphasis, especially for data and marketing scientists, is placed on the analytical techniques and methods to derive insights from data, with reporting and storytelling taking a backseat. Bar charts and pie charts dominate reports, reflecting a tradition that emphasizes familiar formats over optimal clarity. This lack of modernization in data visualization can mask critical insights, degrading the effort put into the analysis.

This is increasingly challenging as audiences grow more diverse—from data-savvy analysts to non-technical clients, with insights driving high-stakes business decisions. Effective data visualization must balance clarity, accuracy, and accessibility while minimizing cognitive load for viewers. Ultimately, visualizations should stand alone, enhancing the narrative to drive decision-making, and bring the advanced analytics often left behind the scenes to the forefront.

Here, we aim to level the playing field by emphasizing storytelling and reporting by synthesizing foundational research from graphical perception, cognitive science, and design principles to inform modern visualization practices, and bring them to life via practical before-and-after examples. We conclude with discussion of how to implement these practices in real-world market research settings given typical constraints like software limitations, team skillsets, and stakeholder expectations.

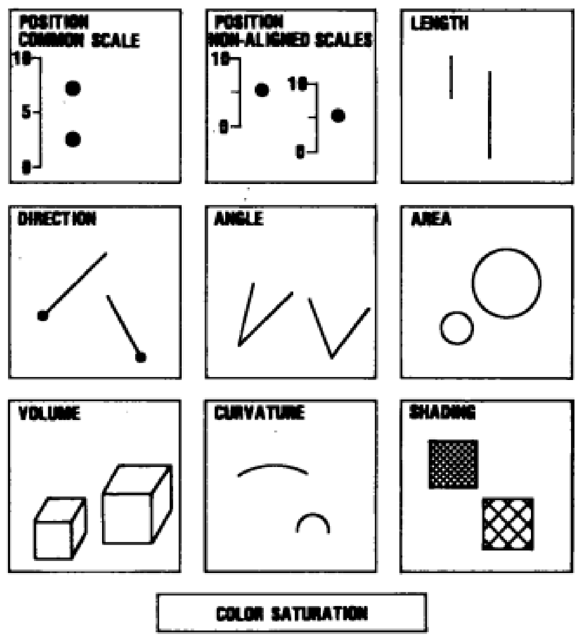
By grounding visualization choices in decades of research (from graphical perception experiments to cognitive load theory and accessibility guidelines) market researchers can communicate data more effectively. Visualization is not merely an aesthetic exercise; it is about optimizing visual cognition. As we will show, principles from Tufte's minimalism[1] to Cleveland and McGill's perception rankings[2] to Kosslyn's cognitive design rules[3] all point toward the same outcome: graphics that let the data speak truthfully and clearly.

PRINCIPLES OF EFFECTIVE DATA VISUALIZATION

Graphical Perception and Visual Encodings

A fundamental concept in data visualization is the concept of visual encodings—the mapping of data variables to visual attributes (position, length, color, etc.) that people perceive (Fig. 1). How effectively viewers decode a chart depends heavily on the choice of encoding. Classic studies by Cleveland and McGill[2,4] systematically tested people's accuracy in reading different graphical elements. They produced a ranked list of encoding effectiveness, finding that position along a common scale is the most accurate for quantitative judgments, followed by position on nonaligned scales, length, angle/slope, area, and finally color saturation or hue. In other words, a dot plot or aligned bar chart (position/length) enables more precise comparisons than a pie chart (angles) or bubble chart (areas), which in turn outperform judgments based purely on color intensity. This has direct implications: for example, when comparing values across categories, aligned bars or points are usually better choices than slices of a pie. Notably, Cleveland and McGill also highlighted that color (hue and saturation) was among the least accurate encodings for judging magnitudes and aside from the perceptual precision issue, heavy reliance on color can introduce accessibility problems for colorblind readers (discussed further below).

Figure 1: Some common visual encodings tested by Cleveland and McGill[2] in seminal work on graphical perception.



Encodings can be grouped into one of three core categories—spatial position/placement, size or direction, and shape/type—echoing earlier frameworks[7,8] that set the stage for systematic visualization design. However, one of the key lessons from these frameworks is that choosing the right encoding for the task is critical: leverage encodings that the human visual system can decode easily and accurately. Position and length should be preferred for precise comparisons, whereas use of area or color to represent quantities should be done with caution (or for secondary variables), since our brains are not as good at discerning fine differences in those channels.

Studies focusing on specific chart types further illustrate these principles. For example, part-to-whole comparisons (such as market share by segment) are often shown with pie or donut charts, yet research consistently finds that such angle-based encodings are less effective for comparing values than bar charts[10] (even when optimized to be as clear as possible[3]). The point is that one must match the visualization type to the purpose: use pie or donut charts sparingly for high-level proportion overviews and prefer more precise encodings when exact value comparison or ordering is important.

In summary, decades of graphical perception research provide a clear direction for modernizing chart choices. We should favor encodings that play to human strengths (e.g., leveraging excellent spatial judgment of aligned lengths) and be wary of those that do not. These insights form the backbone of effective visualization, ensuring that our graphs communicate the data truthfully (no accidental distortions) and efficiently (viewers can decode values with minimal effort).

Cognitive Load and Working Memory

Good visualizations not only use effective encodings but also respect human cognitive limitations. One key consideration is cognitive load, or the total amount of mental effort required to interpret a visual. A core guideline is to limit unnecessary load: include no more (and no less) information than the audience needs. Human working memory is limited—research on visual working memory suggests we can hold only about three to four distinct items or chunks of visual information at once[11]. Researchers have emphasized this limitation in the context of graphics, noting that working memory constraints mean that graphs should be designed to minimize extraneous search and comparison[12]. For example, if a reader is forced to constantly look back and forth between a legend and the chart to match colors with categories (a common issue in complex legends), they are using up precious mental resources on a task that could be alleviated by directly labeling the chart. Each additional cognitive step—deciphering a confusing axis, remembering what a color signifies—adds to the load and risks losing part of the audience. Cognitive load can be taxed further by superfluous content, often referred to as "chartjunk," a concept we'll discuss in detail later.

Cognitive science also highlights the role of prior knowledge and context in interpreting visuals. A viewer's ability to form a correct mental model of a graph depends on their familiarity with the domain[12]. Knowing your audience is essential. This means when designing visualizations, one should consider the expected literacy of the audience: use simpler representations and introduce new or advanced chart types only when they add clear value. It also means providing sufficient contextual cues so that viewers can connect the visualization to their existing knowledge. For example, if showing an index or a scaled value, providing a baseline or reference marker can help ground the interpretation. This is particularly challenging in market research, which often relies on a single set of reporting deliverables that will eventually be consumed by a variety of viewers.

In summary, cognitive science advises us to streamline visuals: make the important parts obvious, avoid clutter, and don't ask the viewer to hold too many pieces in their mind at once.

Color and Accessibility

Visualizations should be accessible to the widest audience possible, including people with color vision deficiencies and other visual impairments. Approximately 300 million people worldwide have some form of colorblindness[13], the most common forms mean that contrasts between red and green hues—a combination unfortunately common in default charts—may be invisible or confusing to those viewers. Therefore, a visualization that uses red and green as primary indicators could easily exclude a portion of the audience from correctly reading the data. Research and guidelines strongly advise to avoid problematic color combinations (swap out red and green for blue and orange) and to use redundant cues in addition to color when distinguishing categories[14].

From a perceptual standpoint, color hue is an inherently weaker channel for precise quantitative interpretation, but it is very useful for categorical differentiation and for drawing attention. Colin Ware's work on visual perception[15] describes how color can pre-attentively signal groupings or outliers if used thoughtfully, but also how it can mislead if overused or if the color scheme implies an order that doesn't exist. Thus, if using a heatmap or colored scale, it's best to use a well-chosen, perceptually uniform colormap[16] and not to rely on subtle color differences to convey critical information.

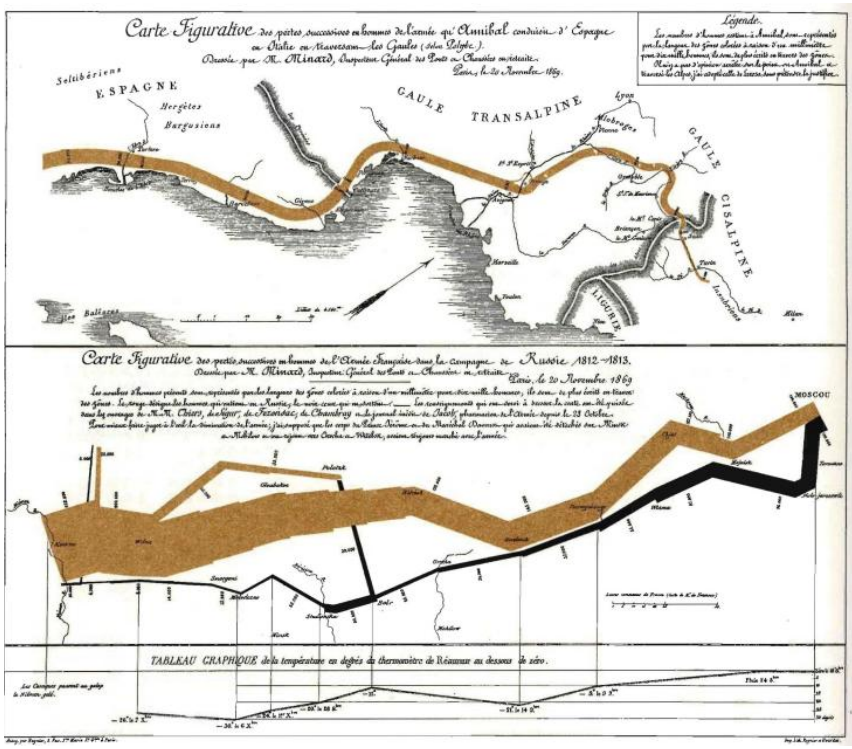
In practical terms, achieving accessibility might mean using color-blind safe palettes, using texture or shape differentiation for overlapping areas in a chart, maintaining appropriate contrast ratios[17], and always labeling data directly instead of relying solely on color-coded legends. The effort put into making a visualization accessible can dramatically improve its effectiveness for all users, not just those with known impairments, because it generally leads to cleaner, higher-contrast, and well-annotated graphics. In short, accessibility improvements are usability improvements.

Clarity, Honesty, and Minimalism in Design

Edward Tufte's principle of "maximizing data-ink ratio" epitomizes the drive for clarity and honesty in visualization design. Tufte argued that a good graphical display should devote as much ink (or pixels) as possible to actual data, and minimize the ink used for anything else[1]. This philosophy translates to eliminating unnecessary embellishments, sometimes derisively called chartjunk. Chartjunk includes decorative backgrounds, excessive grid lines, 3D effects, pictorial icons that do not communicate additional information, or any visual element that does not tell the core data story. Tufte introduced two famous rules: erase non-data ink and erase redundant data ink. By removing extraneous elements, we de-clutter the visual, allowing the data to stand out clearly. The goal is not art for art's sake, but a kind of elegant simplicity that enhances comprehension.

One of the most celebrated examples in the history of data visualization is Charles Minard's 1869 chart depicting Napoleon's Russian campaign (Figure 2). Popularized by Tufte[1], this chart exemplifies many of his foundational principles: particularly clarity, honesty, and minimalism. It visualizes six variables (troop size, direction, geography, temperature, time, and distance) without resorting to unnecessary decorative elements, thereby maximizing what Tufte calls the "data-ink ratio." Every graphical component serves a purpose in conveying the narrative, making the visualization both analytically rich and resonant.

Figure 2: Minard's 1869 visualization of Napoleon's campaign, as reproduced by Tufte[1].



Though the context is historical, the lesson remains urgent for modern market research: visualizations must not only display data, but do so in a way that enhances comprehension, avoids distortion, and respects the cognitive limitations of viewers. The elegance and effectiveness of Minard's work underscore the value of visual discipline and intentionality.

That said, simplicity should not come at the cost of information or accuracy. Some visuals benefit from contextual cues like reference lines or annotations—these are not chartjunk if they help interpret the data. One way to ensure honesty and clarity is to follow cognitive design principles (most famously outlined by Kosslyn[6]). Both Tufte and Kosslyn, along with other experts, stress integrity in data visualization. This includes using consistent scales, zero-baselines for bar charts when applicable, and proportional representations. These guidelines ensure that the visualization is honest: it does not inadvertently deceive. Sometimes, well-intentioned designers violate these rules under the pressure to make charts look interesting.

In practical market research work, clarity and honesty might manifest in choices like using direct labels instead of legends, sorting bars in a meaningful order so that the ranking is evident and providing context in the chart to avoid isolated numbers that invite misinterpretation. It also means being cautious with interactive or "fancy" visualizations that might hide data. By adhering to minimalism and cognitive principles, we aim for visuals that are straightforward, truthful, and tailored to how people perceive and think.

Having reviewed these key principles from literature—graphical perception (choosing effective encodings), cognitive load minimization, accessibility, and clarity/honesty through minimalist design—we now turn to concrete examples.

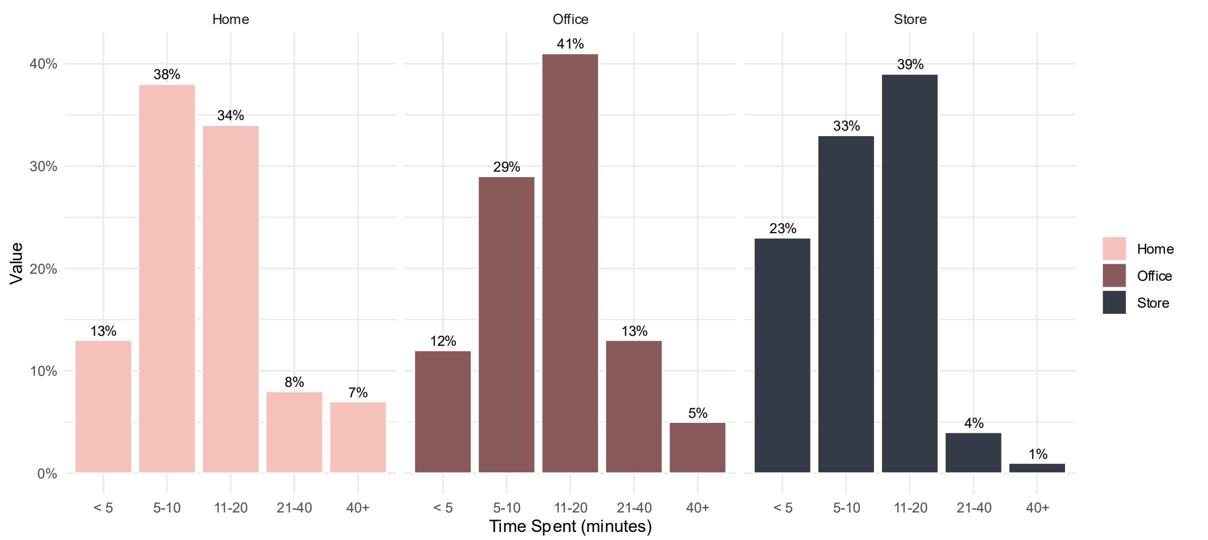
Examples in Practice

In this section, we present four illustrative examples of improving data visualizations commonly encountered in market research. Each example will describe an initial "before" visualization that has issues (such as misleading design, clutter, or suboptimal encoding), then detail the "after" state—a redesigned visualization that applies core principles to better communicate the insights.

Example 1: Handling Unequal Bin Widths in Bar Charts

Before: A bar chart with unequal bin widths on the x-axis (Fig. 3) remains a subtle but consequential visualization flaw. This example depicts a time-use survey where respondents reported how many hours they spend in each setting—Home, Office, or Store—using ranges like "<5," "5–10," "11–20," "21–40," and "40+." At first glance, the chart may appear conventional and well-structured, with bars indicating the percentage of respondents within each range for each setting. However, there is a core issue: the bins are not of equal width in the variable they represent yet are plotted with equal visual width.

Figure 3. Before visualization of a bar chart with unequal bin widths, treating bins as equal —e.g., income ranges with unequal spans plotted as equal-width bars.



This mismatch distorts the interpretation of the data. For example, the "11–20" bin spans 10 hours, while the "40+" bin is open-ended and could include a vast range of values. Treating these categories as equal-width bars implies that each bin represents an equal unit of measurement. As a result, visual comparisons of bar height suggest differences in density per hour range, but the visual encoding doesn't adjust for the varying widths of those ranges. A tall bar for a wide bin may appear more significant than it is, simply because it aggregates responses over a broader interval.

In the Office group, for instance, the "11–20" hour bin shows the highest percentage of respondents, but this category covers more than twice the span of "5–10" and is significantly less than the "21–40" span. If viewers interpret bar height as directly proportional to time-based density, they may overestimate how concentrated respondents are in that middle range.

This figure underscores the need for caution when visualizing quantitative ranges as bar charts. Unless bin widths are equal or density is explicitly normalized, such charts can lead viewers to flawed conclusions—emphasizing total counts where per-unit comparisons would be more appropriate. In cases involving time, income, or other continuous metrics, unequal bin widths should be carefully considered and clearly conveyed to avoid misleading visual inferences.

After: To remedy the misleading effects of unequal bin widths, the revised visualizations (Fig. 4a and Fig. 4b) demonstrate two effective alternatives that ensure the chart accurately reflects the data distribution. Each addresses the core problem—misinterpreting frequency due to unequal interval spans—by using visual encodings that align bar area or size with the variable's true range.

Figure 4a uses a proportional-width bar chart in which each bar's width corresponds directly to the span of the bin it represents. Wider time intervals, such as "40+," are shown with a broader base, while narrower ones like "<5" are thinner. The bar heights represent density: the proportion of respondents per unit of time. This format corrects the visual imbalance present in the original chart, where bars of equal width falsely implied equal interval sizes. In the corrected version, if the "11–20" bin is taller than "5–10" but also twice as wide, the viewer can now better judge how much of the apparent frequency difference is due to bin span rather than actual response density. This approach is faithful to histogram principles: bar area now reflects the underlying data count.

Figure 4a: One alternative to Fig. 3—a proportional-width bar chart.
Note that here, some assumptions must be made about width
(e.g., How wide do we make an un-bounded bin?).
Pointing this out to viewers is key for transparency.

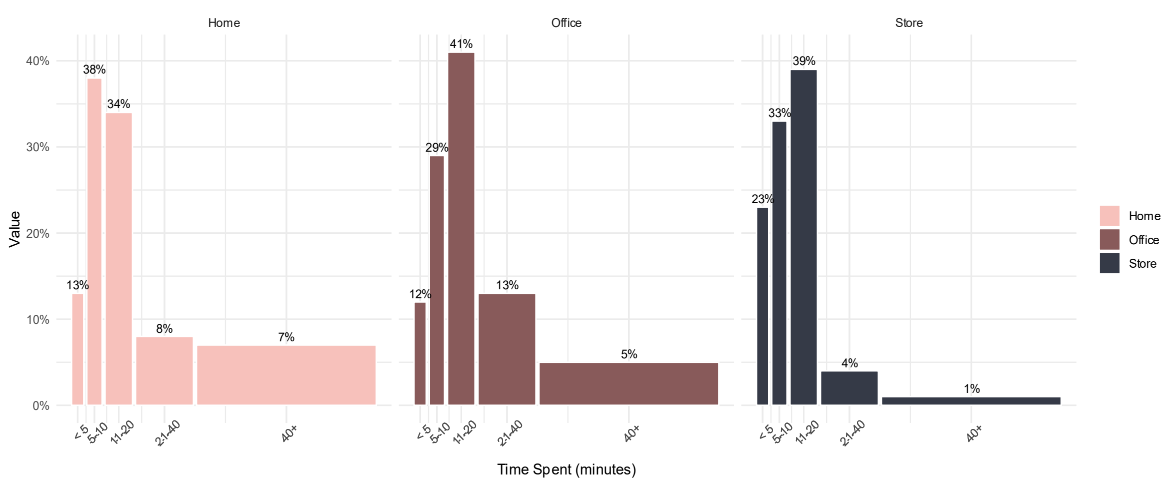
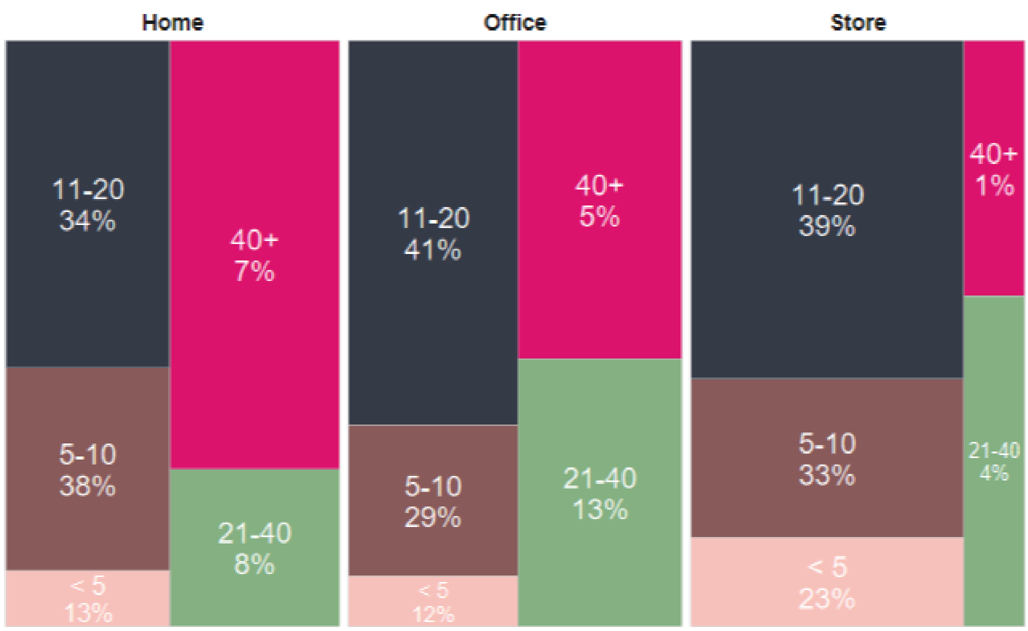


Figure 4b presents an alternative using a treemap. Here, each rectangle's area represents the product of the bin size and the proportion of respondents in each time-use bin. This eliminates the issue of implied equality in axis spacing altogether. Because treemaps are not structured on a continuous axis, they avoid the temptation to interpret spacing or bar height as density per unit interval. The visual emphasis is placed instead on total share which is ideal if the goal is to communicate aggregate composition rather than fine-grained distribution. Treemaps work especially well when the bin categories are conceptually distinct or when the range spans are irregular and non-numeric.

Treemaps are not a panacea, and we recommend an abundance of caution when presenting them to unfamiliar viewers, as well as adding explicit annotations indicating that the area of the rectangle represents the combination of the bin size and the respondent proportion. Effective use will depend on the goal. If the goal is to show the distribution of respondent density across a numeric axis, a proportional-width bar chart is the best choice (Fig 4a; e.g., most respondents spend a moderate amount of time in each setting, with no large differences among the groups). Instead, if a viewer is interested in aggregate composition that takes into account accurate part-to-whole calculations (e.g., we can see a large shift towards longer times in the store compared to office or home, despite high proportions in narrow bins), then a treemap may be a more appropriate choice.

Figure 4b: Another alternative to Fig. 3—a treemap.
The goal of the treemap is to compare total bin frequencies and are most effective in situations where their bins are very unequal or represent un-bounded ranges. Note: the area of each rectangle is mapped to the size of a given bin and the proportion of respondents in a given bin (the percentage label).



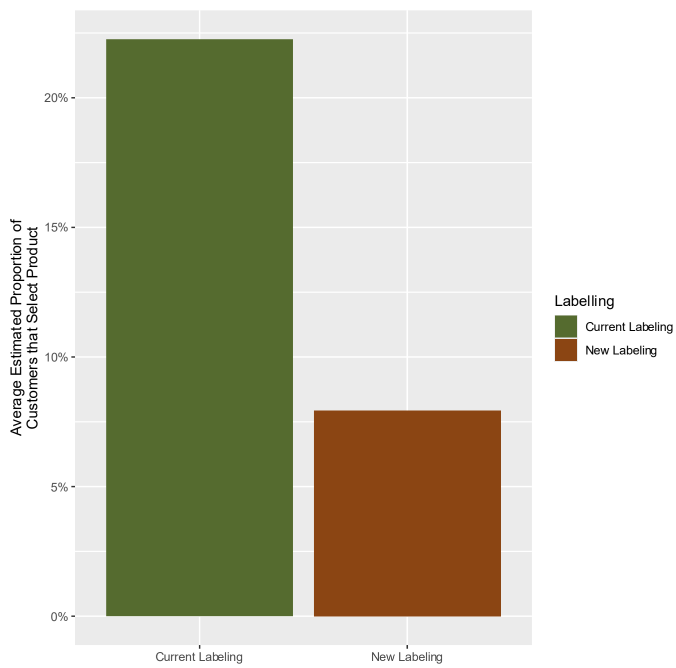
Both methods achieve the key aim: they make the underlying bin widths visually transparent. Which method is preferable depends on the communication goal. If you want to show respondent density across a continuous variable, the proportional-width histogram is ideal. If you want to compare total bin frequencies in a visually compact space without implying continuity, the treemap is elegant and effective.

In practice, these designs can be created in tools like the ggplot2 R package[19] by manually setting bar widths or calculating density per unit for the y-axis. Treemaps can be generated with libraries like treemapify in R or plotly and matplotlib in Python[20,21]. Regardless of the method, the principle is clear: when visualizing binned data, ensure that visual area, not just bar height, aligns with bin span and frequency.

Example 2: Replacing Bar Graphs of Averages with Distributions

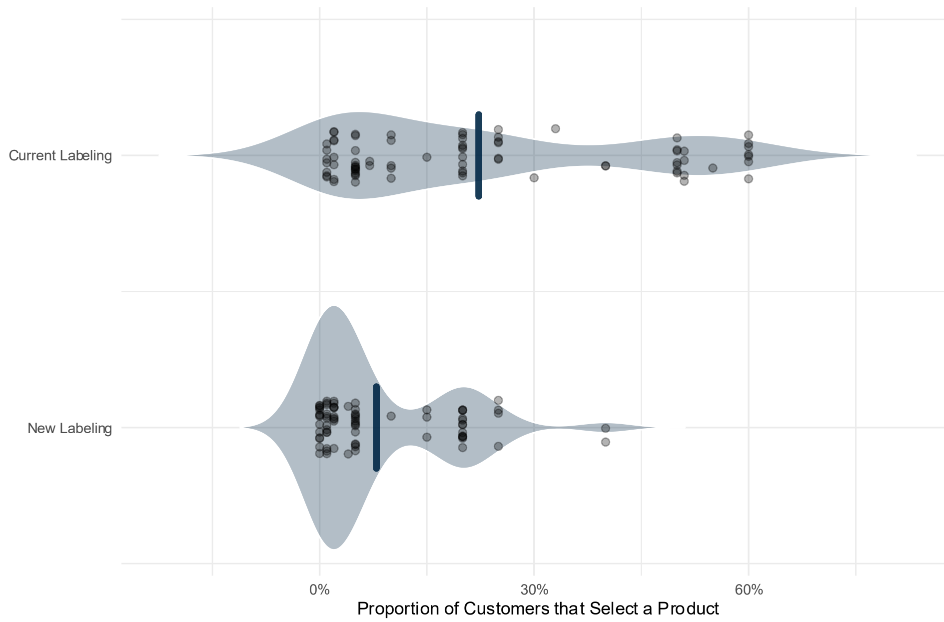
Before: A common scenario in reporting is to summarize a continuous metric with a single aggregate statistic such as a mean. Often, these are presented as bar graphs of the mean for different conditions or time points, sometimes with error bars. While this approach is pervasive, it can be problematic. This example (Fig. 5) shows data from a survey question asking participants to estimate a certain percentage.

Figure 5: A simple (and problematic) bar chart showing a comparison of mean of two groups.



If we only show the average of the estimates before and after, we lose crucial information about the distribution and individual differences. Two groups of respondents could have very different distributions of answers and yet end up with the same mean. In fact, many different data distributions can lead to the same bar graph, meaning that the bar may mask whether the data were tightly clustered, bimodal, widely varied, etc. This is a recognized issue in scientific visualization and work has shown that bar charts of means often conceal important data patterns and researchers have advocated for more informative plots that show all data points or distributions[22].

Figure 6: An updated mixed violin plot showing the same data as in Figure 5.
Here, each dot represents a given respondent, and dark vertical bars represent the mean of each group. Points are jittered vertically for clarity and separation and are plotted with transparency to more clearly indicate data density in particular areas. Violins highlight the regions of highest density for each distribution.



Statistical annotations like error bars help, but the viewer cannot see the shape of the data or the paired nature. The cognitive danger is that a bar implies a uniform entity from zero up to the height of the mean, which is a misleading representation of the raw data.

After: The improved approach is to display the distribution of data or at least individual data points, especially for small to moderate sample sizes. There are several options to replace the bar chart of means:

1. A strip plot (dot plot) or scatter plot of individual responses for each condition, potentially with lines connecting paired observations.
2. A box plot or violin plot, which shows median, quartiles, and overall distribution shape. This gives a sense of spread and any asymmetry or outliers.
3. A before-after slope graph, which is essentially plotting the value for each respondent at Time 1 vs. Time 2 and drawing a line between them. This directly shows how each respondent moved (e.g., most lines slanting downward if most estimates decreased).
4. If the sample size is large and individual points would clutter, a density plot or histogram for each condition can work, possibly overlaying them or placing them side by side, to show how the distribution changed.

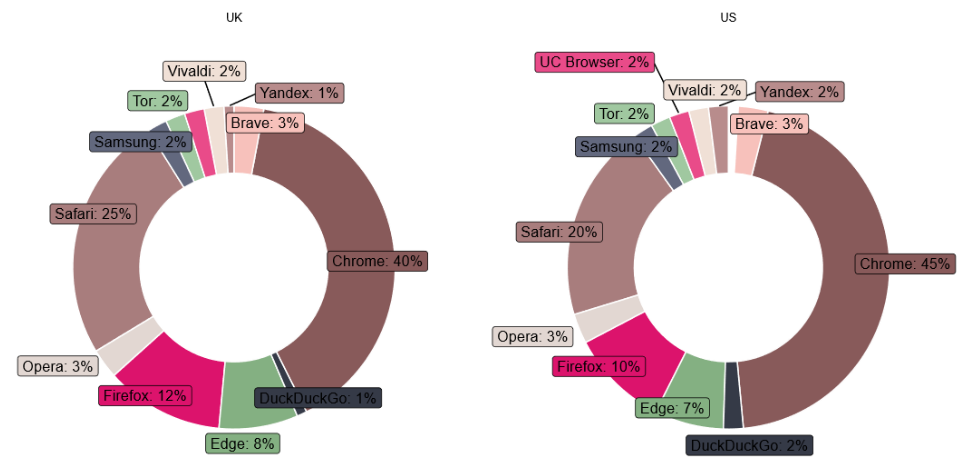
We opt for the mixed violin plot (2, above), as it's the simplest example and conveys the same core message as the original plot (before and after are, on average, different; Fig. 5).

By moving from bars of averages to plots that display distributions, we adhere to Tufte's directive of maximizing data ink in a meaningful way. Here, the "data ink" is every respondent's answer. There are many occasions where showing individual responses, or at least the distribution, is both possible and beneficial, particularly in situations when surveys allow for numeric input or when non-numeric inputs generate numeric values (e.g., individual utilities in conjoint). The benefit is an honest and richer view of the data, which can prevent erroneous interpretations that might arise from just comparing means.

Example 3: Improving Part-to-Whole Comparisons (Moving Beyond Pie Charts)

Pie charts and donut charts are a go-to for showing parts of a whole, but rank low in perceptual accuracy for comparing magnitudes. Human eyes are not very good at comparing angles or areas that are separated spatially[23,24]. Additionally, pie charts struggle when the goal is to rank or precisely compare values: if a stakeholder asks, "Which web browser is first in a given market and how much ahead is it compared the browser ranked second?," a pie chart is not well suited—a bar chart or sorted list would be more straightforward.

Figure 7: A problematic, two-panel donut chart showing browser usage across two groups (the US and the UK).
Core issues include: poor color contrast ratios in labels, difficult comparisons across groups and the inability to quickly distinguish differences among items that occupy small proportions of the total.

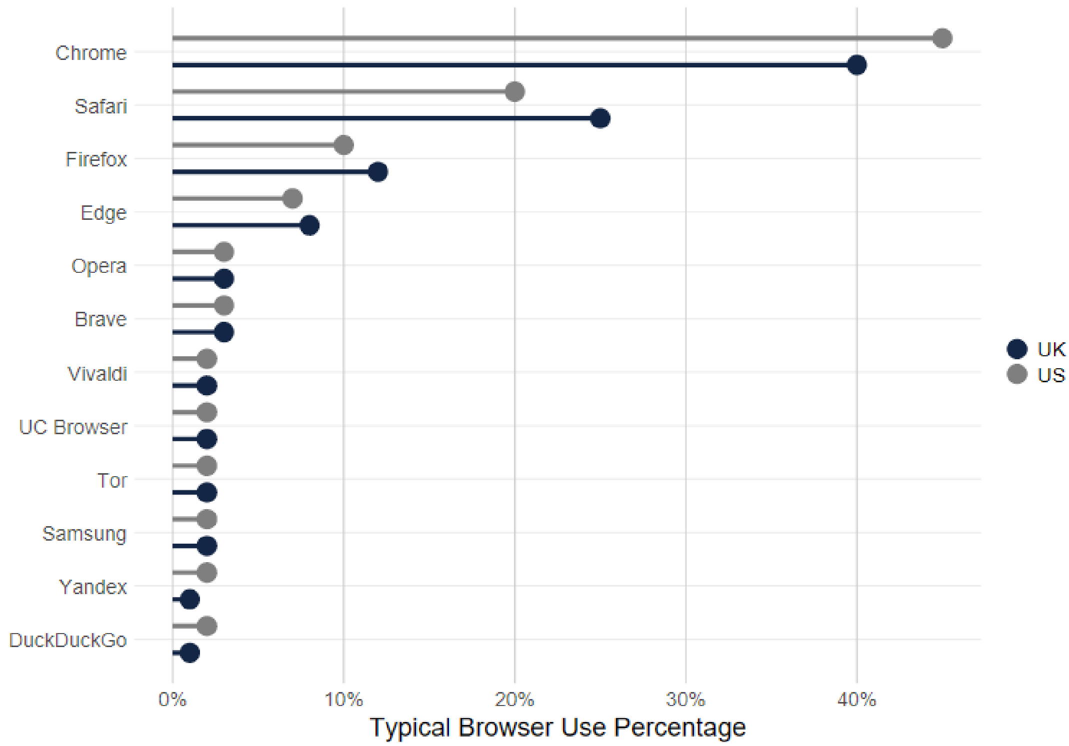


Research has repeatedly shown that position along a common scale (as in a bar chart) yields faster and more accurate comparisons than angle or area in pies[2,4,23]. It's not that pies convey no information, in fact, they can be can faster than a table of numbers for getting a rough sense of proportions, but are suboptimal when finer discrimination is needed.

Before: Here, we show an example illustrating the percentage of survey respondents using different web browsers, with slices for Chrome, Safari, Firefox, Edge, and other options. If comparing two segments or markets, sometimes two pies are shown side by side (Fig. 7). Often, the goal is to compare slices within or across charts: if Firefox is 15% in one pie and 10% in another, one slice might be slightly bigger than the other—but unless they are adjacent, it's difficult to tell the difference. With many categories, pies also get cluttered with tiny slices and labels, often needing color codes and legends that strain attention and working memory.

After: To address the shortcomings of pie and donut charts in part-to-whole comparisons, especially when precise comparisons or rankings are needed, a lollipop chart offers a cleaner and more cognitively efficient alternative. Rather than relying on angles or areas, this approach uses aligned positions along a common horizontal scale-making it easier to assess both the relative order and the magnitude of differences between categories.

Figure 8: A redesigned lollipop chart comparing browser usage in the US and UK. The horizontal positioning enables accurate ranking and value estimation across categories and between markets. If distinguishing between categories with similar percentages was a goal, breaks or faceting (multi-panels) could also be used.

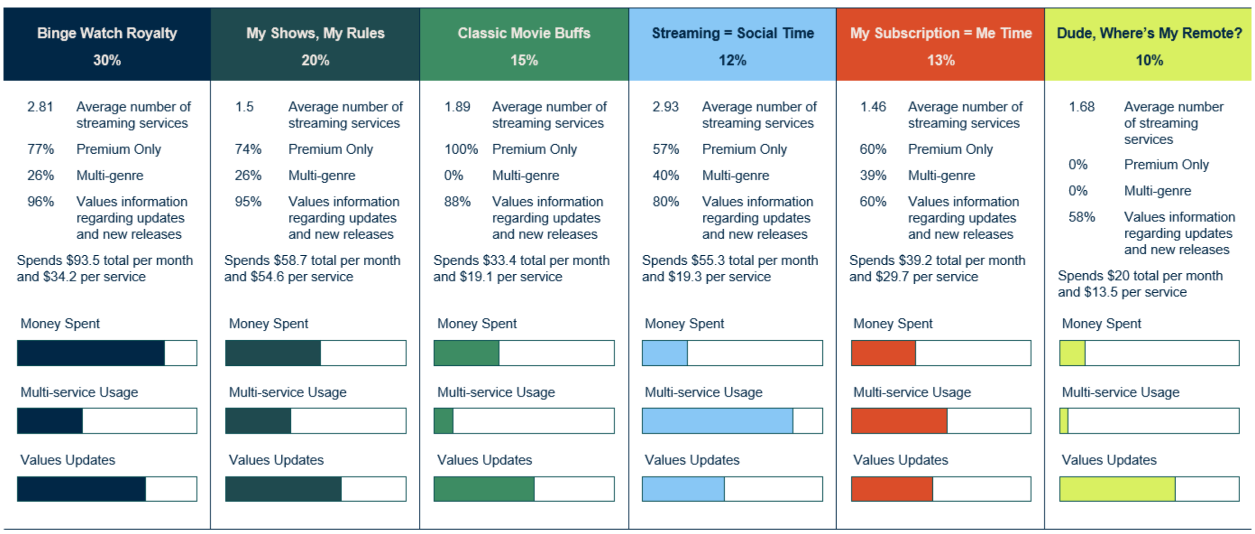


Each browser is listed on the vertical axis, with its usage rate indicated by the horizontal position of a circular marker connected by light stems for added visual anchoring. This format supports quick, intuitive comparisons both within each market and across them. By sorting the categories and standardizing the horizontal scale, stakeholders can easily answer questions such as “Which browser is third in the UK?” or “How much higher is Chrome usage in the US than in the UK?” without mentally estimating slice sizes or decoding color-coded legends. The visual simplicity also scales better with more categories or more granular segmentation.

Example 4: Visualizing Multi-Dimensional Segment Profiles Clearly

Segmentation is a foundational practice in market research—dividing audiences into meaningful groups based on shared behaviors, attitudes, or preferences. Each segment typically carries a distinct profile spanning multiple variables, such as usage intensity, preferences, and value perceptions. The challenge isn't generating these data—it's presenting it in a way that decision-makers can quickly grasp and compare. When a segmentation scheme includes five or more dimensions, even well-structured visuals can fall short if they overwhelm rather than clarify.

Figure 9: A detailed comparison table summarizing six streaming customer segments across multiple behavioral and attitudinal metrics. While precise, the format requires viewers to make manual cross-column comparisons and interpret bar-fill graphics without a unified scale.

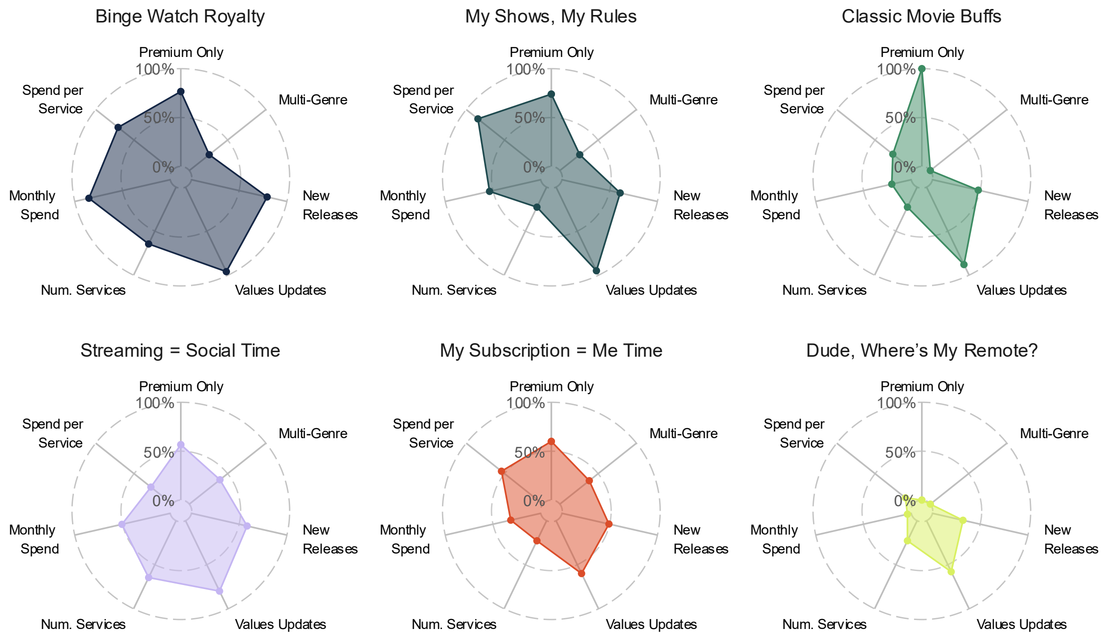


Before: A common approach to segment profiling is the summary table or slide format (Fig. 9). Each segment is displayed in its own column, with key metrics stacked vertically: e.g., number of services, percent premium, percent multi-genre, update value. Visually, the structure is clean, and quantitative precision is high. But cognitively, it places a heavy burden on the viewer: comparisons require scanning across columns for each row, holding multiple values in working memory, and mentally calculating which segment leads on a particular attribute.

In cases where segment names are long and attribute counts grow, this format quickly becomes a wall of text and blocks, more suitable for reference than insight. Though it preserves detail, it sacrifices synthesis.

After: To overcome the limitations of linear tables, radar charts offer an integrated, shape-based alternative that turns segment profiles into visually distinct patterns (Fig. 10). Here, each axis represents a normalized metric such as spending per month, multi-service usage, or interest in content updates—and each segment is plotted as a colored polygon that connects its values across all axes.

Figure 10: A radar chart showing the same six streaming segments across seven standardized metrics.
Each polygon's shape captures the relative strengths and weaknesses of that segment, supporting fast and intuitive profile comparison across groups and dimensions.



This format allows the viewer to instantly perceive differences in profile shape and relative standing. Instead of mentally scanning rows and columns, stakeholders can spot which group peaks on "Spending" while dipping on "Services Used," or which segment hugs the center, indicating consistently low engagement across traits. The contours become visual signatures of behavior. Importantly, all segments share a common scale and orientation, so comparisons are intuitive. This approach transforms segmentation profiles from lists of statistics into cohesive visual stories. It's a powerful upgrade when you want to synthesize multidimensional differences into a single, digestible view.

Balancing Best Practices with Real-World Constraints

The above examples and principles make a compelling case for modernizing data visualization in market research. However, implementing these best practices in the real world involves navigating practical constraints: legacy tool limitations, tight timelines, stakeholder preferences, and varying levels of data visualization literacy among both clients and colleagues.

Tooling: GUI vs. Code and Software Limitations

One significant factor is the tools and technologies used in day-to-day work. Many market researchers rely on GUI-based tools like Excel, PowerPoint, SPSS, or Tableau for their charting needs. These tools make it easy to produce standard bar and pie charts quickly, but they may not support advanced or custom visuals (or if they do, the features might be hidden or require cumbersome manual steps). For example, creating a radar plot with custom colors and transparencies might be straightforward in R or Python with a few lines of code, but is very cumbersome in Excel or PowerPoint. This leads to a status quo bias: people stick to what's readily available in the software they know, even if it's not the optimal way to communicate the data.

Code-based tools (like R's ggplot2[19], Python's matplotlib/seaborn[20,21], or JavaScript libraries like D3[26]) offer far more flexibility and power to implement the kinds of visualizations we showcased. They also facilitate reproducibility—once a script is written to produce a certain chart, it can be reused and updated with new data easily. Finally, all of the toolkits above are completely free, and maintained and improved-upon by vibrant contributing communities. However, adopting code-based tools has a learning curve. Not every team member or stakeholder is comfortable with programming, and organizations may not allocate time for employees to develop those skills. Additionally, integrating code outputs back into traditional deliverables is an extra step that some view as a hassle, or potentially disruptive to fast-paced timelines. Overall, there is a trade-off between ease of use, timing, and flexibility.

Strategies to Manage Limitations

- **Incremental tool introduction:** Rather than forcing an entire team to switch to R/Python overnight, introduce code-based visualization by degrees. For instance, start with one or two "power users" who create custom graphs for high-profile projects, demonstrating the impact. Show these alongside the usual charts to build familiarity.
- **Templates and automation:** Develop a library of reusable code templates or functions for common needs (e.g., an R function that takes a dataset and outputs a nicely formatted dot plot with significance stars, etc.). This lowers the barrier for others to use code outputs—they don't need to understand every line, just how to run it with their data.
- **Tool training:** Provide training sessions or resources to get team members up to speed with new tools. Emphasize that learning a bit of code for visualization can save time in the long run (no more hacking Excel to do what it wasn't meant to).
- **Leverage advanced features in familiar tools:** Many are unaware that even Excel can do things like error bars, transparent fills, or scatterplot matrices with some effort. Teaching a few advanced tricks in the tools people already use can produce immediate improvements.
- **AI-powered support:** AI can act as a quick upskill magnifier, allowing those with less experience to quickly and effectively learn how to implement data visualizations in platforms that typically require extensive training, in addition to recommending new visualizations that are grounded in best practices and appropriate for the data at hand. However, it's important to implement careful review because without oversight, AI makes it easy to go down unfruitful paths.

Stakeholder Expectations and Perception Momentum

Another real challenge is stakeholder expectations and resistance to change. Clients and managers often have a fixed notion of what a report should look like. This perception momentum (the inertia of familiar visuals) can make it risky to deviate too far from the norm. There is a psychological comfort in recognizing the format, even if the new format is objectively better.

Managing Expectations

- **Explain and educate:** When presenting a new type of chart, take a moment to explain it. For example, "We used a violin plot here instead of the usual bar chart because it allows you to see the underlying data and patterns in segments of respondents. This gives a more nuanced view than just the average." A brief rationale can turn initial confusion into appreciation.
- **Gradual changes:** Don't overhaul every visualization in a key client report all at once. Pick one or two high impact improvements to introduce in a meeting, gauge reaction, and iterate. Start with internal stakeholders who are more receptive, then extend to clients once the approach is refined.
- **Align with stakeholders' goals:** Emphasize how the improved visualization helps them: "This layout makes it easier for your customers/your boss to grasp the point quickly," or "This chart is more colorblind-friendly, ensuring no one on your team misreads it." When clients see it as enhancing their communication or decision-making, they are more likely to accept it.

Data Literacy and Communication

The level of data literacy among the audience can vary widely. Some senior executives just want the headlines and may not spend time interpreting a complex chart; others might be analytically inclined and appreciate a deeper dive. Modern visualizations often carry more information (showing distributions, multiple facets, etc.), which can be immensely valuable for a data-literate reader but potentially overwhelming for a less savvy reader if not guided. To address this, one must pair visualization with good communication:

- **Descriptive titles and captions:** Each graph should have a clear takeaway in the title (e.g., "Segment A is the primary target for increased spend, while Segment B is most likely to add additional services" rather than just "Segment Metrics"). This way, even if someone isn't fully comfortable with the graph format, the title drives the narrative home.
- **Annotations:** Call out key points on the chart itself with arrows or text. In our examples, we might annotate, "Most respondents decreased their estimate" on the violin plot, or "Edge usage is 7% higher in Market B" on the bar chart. These act as guideposts for readers.
- **Provide interpretation in text:** Always accompany complex visuals with a narrative in the report or presentation. For example, after showing the distribution plot, a sentence like "This chart shows the full range of responses: notice how the distribution narrowed after the intervention, indicating greater agreement among respondents" helps ensure the insight isn't missed.

Adoption Challenges and Strategies for Change

Implementing these modern practices often requires organizational buy-in and cultural change. It's one thing for an analyst to make a better chart; it's another for an entire team to adopt new standards and for clients to request them. Here are some strategies drawn from the presentation and industry experience:

- **Show the business value:** Continuously highlight why these visualization improvements matter. For instance, quicker understanding leads to faster decision-making. Improved accessibility means a broader audience can use the insights. Clarity can reduce miscommunication and the need for follow-up explanations. If possible, collect anecdotes or metrics: did a client react positively ("We finally see what's going on!") or did a project get approved because the data story was clearer? These help justify the effort.
- **Champions and training:** Identify or develop a few data visualization champions within the organization. These individuals stay updated on best practices and tools and disseminate that knowledge. They might host brownbag sessions, share before-and-after makeovers internally, or create an internal "gallery" of effective graphs as references.
- **Establish guidelines or templates:** Formalizing some of these practices into company guidelines can reinforce their use. For example, a guideline might state, "For any quantitative comparison of more than two numbers, avoid pie charts; use bar or dot plots." Providing templates that have these baked in—like a PowerPoint template with pre-set chart styles in the right colors and fonts—can nudge everyone to comply simply by convenience.
- **Client collaboration:** Sometimes clients themselves need to be brought along the journey. If you have a forward-thinking client, you can partner with them to pilot new visualization approaches. Alternatively, if a client is very conservative, you might use more innovative visuals internally to derive insights and then translate them into the simpler formats for delivery, as a compromise. Over time, as trust is built, you can try introducing the improved visuals to them gradually.

In the end, driving change in visualization practice is similar to any change management: it requires demonstrating why the change is beneficial, enabling people to make the change, and making it as easy as possible to do so. One doesn't need to abandon all old charts overnight; rather, steadily introduce improvements that clearly enhance the storytelling power of data.

Functional Recap

To close, here is a summary of key principles that can support practitioners in making more effective and trustworthy data visualization choices:

- **Prioritize perceptual efficiency:** Favor visual encodings that the human visual system processes with the greatest speed and accuracy. Position along a common axis is typically the most effective and should be the default for quantitative comparisons.
- **Use pie and donut charts judiciously:** Reserve these formats for high-level overviews of proportions where precision is less critical. For tasks involving exact comparisons or ranking, opt for bar charts, dot plots, or other encodings with more precise visual cues.
- **Eliminate chartjunk:** Strip away non-informative elements that distract rather than clarify, such as decorative backgrounds, excessive gridlines, 3D effects, and icons that are more ornamental than meaningful.
- **Maintain visual integrity:** Avoid misleading your audience, even unintentionally. Ensure scales are consistent, baselines are present when needed (especially for bar charts), and that graphical elements accurately represent the underlying data.

- **Know your audience:** Tailor visualizations to the intended group of viewers and try to avoid a one-size-fits-all approach. Viewers have different skills and experience they bring to the table—build accordingly.
- **Build for accessibility:** Avoid inadvertently excluding viewers by thinking deeply about color, size and contrast ratios. Modern tools are available to quickly assess palettes and make data-driven decisions—or avoid color altogether and opt for a more powerful visual encoding.

CONCLUSION

The landscape of data visualization is continually advancing, informed by research in human perception, cognition, and design. Modernizing data visualization practices in market research is both a challenge and an opportunity. As we have discussed, embracing principles from the likes of a pantheon of thinkers on data visualization leads to graphics that are clearer, more accurate, and more inclusive. The literature provides a strong foundation: use effective encodings (position and length over area and color), reduce cognitive overload (focus on what matters, respect the 3–4 item working memory limit), design for accessibility (don't let a large proportion of the population be left guessing about color choice), and strive for honesty and simplicity (maximize the data, minimize the junk).

Through the four practical examples, we saw these principles in action. Each "before" visualization—whether a misleading bar chart, an uninformative average, a hard-to-read pie, or an overplotted segment graphic—was improved by applying a mix of common-sense and research-backed strategies: adjust the design to the data, not the data to the design. The "after" versions used the same data to tell a more meaningful story. These examples underscore that improving visualizations doesn't always require exotic new chart types; often, it's about refining or combining basic charts in smarter ways and paying attention to details (like bin widths, ordering or migration to different encodings).

Finally, we addressed the critical bridge between theory and practice. Change in how we visualize data must consider tools, habits, and audience. By starting small, educating stakeholders, and demonstrating value, research teams can gradually raise the bar for data visualization in their deliverables. The benefit is not academic—it is practical impact. Insights that would have remained hidden become apparent. Audiences that might have been confused become engaged. Decisions that might have been rushed or ill-informed can be made with confidence.

In the evolving world of analytics and insights, those who communicate data effectively have a distinct advantage. As data scientists and analysts, we often focus primarily on analytics and methods and not communication—if our audience can't understand the insight, all the technical work is in vain. Modern visualizations can set agencies apart, showing clients that you not only have the data, but you can make them understand it. The future of market research visualization is one that blends the scientific rigor of perception research with the art of storytelling. By modernizing our practices, we honor both the data and the audience.

In conclusion, modernizing data visualization is not about being trendy—it's about being clear, truthful, and impactful. By implementing the principles and practices outlined in this paper, market researchers can ensure their work not only informs but resonates. The data we work so hard to collect and analyze deserves to be seen in the best light; it's our job as insight professionals to shine that light clearly and brightly. The tools and knowledge are at our disposal—now it's about making the commitment to use them, one chart at a time.

J. Keaton Wilson

Ben Cortese

CITATIONS

1. Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.

2. Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association, 79*(387), 531–554. https://doi.org/10.2307/2288400

3. Kosslyn, S. M. (2006). *Graph design for the eye and mind*. Oxford University Press.

4. Cleveland, W. S., & McGill, R. (1987). Graphical perception: The visual decoding of quantitative information on graphical displays of data. *Journal of the Royal Statistical Society: Series A (General), 150*(3), 192–229. https://doi.org/10.2307/2981473

5. Heer, J., & Bostock, M. (2010). Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 203–212. https://doi.org/10.1145/1753326.1753357

6. Zeng, Z., & Battle, L. (2023). A review and collation of graphical perception knowledge for visualization recommendation. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Article 820, 1–16. https://doi.org/10.1145/3544548.3581349

7. Bertin, J. (1983). Semiology of graphics: Diagrams, networks, maps (W. J. Berg, Trans.). Madison, WI: University of Wisconsin Press.

8. Mackinlay, J. D. (1986). Automating the design of graphical presentations of relational information. ACM Transactions on Graphics, 5(2), 110–141. https://doi.org/10.1145/22949.22950

9. Skau, D., & Kosara, R. (2016). Arcs, angles, or areas: Individual data encodings in pie and donut charts. *Computer Graphics Forum*, *35*(3), 121–130. https://doi.org/10.1111/cgf.12888

10. Li, Y., Berger, E. D., Kahng, M., & Xiong Bearfield, C. (2024). *From perception to decision: Assessing the role of chart types affordances in high-level decision tasks*. arXiv preprint arXiv:2410.04686. https://doi.org/10.48550/arXiv.2410.04686

11. Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science*, *15*(2), 106–111. https://doi.org/10.1111/j.0963-7214.2004.01502006.x

12. Hegarty, M. (2011). The cognitive science of visual-spatial displays: Implications for design. *Topics in Cognitive Science*, *3*(3), 446–474. https://doi.org/10.1111/j.1756-8765.2011.01150.x

13. Birch, J. (2012). Worldwide prevalence of red-green color deficiency. *Journal of the Optical Society of America A*, *29*(3), 313–320. https://doi.org/10.1364/JOSAA.29.000313

14. Kilin, I. *Data visualization for colorblind readers*. Datylon Blog. Retrieved June 2, 2025, from https://www.datylon.com/blog/data-visualization-for-colorblind-readers

15. Ware, C. (2012). *Information visualization: Perception for design* (3rd ed.). Boston, MA: Morgan Kaufmann.

16. Harrower, M., & Brewer, C. A. (2003). *ColorBrewer.org: An online tool for selecting colour schemes for maps*. *The Cartographic Journal*, *40*(1), 27–37. https://doi.org/10.1179/000870403235002042

17. World Wide Web Consortium (W3C). (2023). *Web Content Accessibility Guidelines (WCAG) 2.2*. https://www.w3.org/TR/WCAG22/

18. Źróbek, P. (2023, September 24). *Guarding against misleading data*. Medium. https://medium.com/number-around-us/guarding-against-misleading-data-503424ecd457

19. Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2nd ed.). Springer. https://doi.org/10.1007/978-3-319-24277-4

20. Plotly Technologies Inc. (2015). *Collaborative data science*. Montréal, QC: Plotly Technologies Inc. https://plotly.com/python/

21. Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment*. *Computing in Science & Engineering*, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

22. Weissgerber, T. L., Milic, N. M., Winham, S. J., & Garovic, V. D. (2015). Beyond bar and line graphs: Time for a new data presentation paradigm. *PLOS Biology*, *13*(4), e1002128. https://doi.org/10.1371/journal.pbio.1002128

23. Siirtola, H. (2019). The Cost of Pie Charts. In Proceedings of the 23rd International Conference on Information Visualisation (IV 2019) (pp. 151–156). IEEE. https://doi.org/10.1109/IV.2019.00034

24. Siirtola, H., Räihä, K.-J., Istance, H., & Spakov, O. (2019). Dissecting Pie Charts. In D. Lamas, F. Loizides, G. N. Moridis, & L. Nacke (Eds.), Human-Computer Interaction—INTERACT 2019 (Vol. 11747, pp. 688–698). Springer. https://doi.org/10.1007/978-3-030-29384-0_41

25. Li, Y., Berger, E. D., Kahng, M., & Xiong Bearfield, C. (2024). *From perception to decision: Assessing the role of chart types affordances in high-level decision tasks*. arXiv preprint arXiv:2410.04686. https://arxiv.org/abs/2410.04686

26. Bostock, M., Ogievetsky, V., & Heer, J. (2011). *D³ Data-Driven Documents*. *IEEE Transactions on Visualization and Computer Graphics*, 17(12), 2301–2309. https://doi.org/10.1109/TVCG.2011.185

TURNING IT TO 11: A PRACTITIONER-LED COMPARISON OF VOLUMETRIC CONJOINT ANALYSIS TECHNIQUES

DEAN TINDALL
SAWTOOTH SOFTWARE
CHRIS MOORE
MANJULA BHUDIYA
IPSOS

INTRODUCTION

Volumetric conjoint analysis is increasingly seen as an important evolution in discrete choice modelling, particularly for clients and stakeholders who seek to uncover more information about the volumes they are likely to sell, instead of, or in addition to the traditional preference shares. While traditional Choice-Based Conjoint (CBC) does a fantastic job of identifying what people are most likely to choose, volumetric conjoint goes further—it quantifies how much of a product or service respondents are likely to consume. This additional layer can allow analysts to better simulate market demand, forecast volume impacts of pricing or promotional strategies, and guide operational or inventory planning. In short, it more closely links preference shares with realistic commercial outcomes.

Despite this, volumetric conjoint remains underutilized in the research industry. Many practitioners struggle with implementation—not due to lack of interest, but because of the increased complexity (perceived and actual) in both survey design and analytical execution. Issues such as open-ended quantity input, variability in response realism, and the challenge of simulating None selections have created a barrier to widespread adoption. Modelling volumetric responses often demands more advanced methods and potentially time-consuming analysis than standard CBC approaches, making it less accessible to analytical teams without specialised technical expertise.

This paper presents findings from a practitioner-led study designed to bridge this gap between aspiration and application. Five candidate methods are compared for modelling volumetric conjoint data, evaluating them not only on statistical performance but also on usability, scalability, and their alignment with real-world commercial needs. The intent is not to declare a definitive "best" method, but to offer clear guidance for practitioners navigating this increasingly important space. For this paper the focus was on methods that are more easily implemented using standard, widely available software which will be familiar to many practitioners.

To do so, a mid-complexity conjoint study was designed, centred on the potato crisps category in the UK. This familiar, competitive, and promotion-sensitive category allows exploration of how different models handle volume decisions under a range of product configurations. Throughout this paper, the aim is to provide a balance of technical depth and practical interpretation, hopefully giving practitioners the confidence to undertake these techniques.

Background and Objectives

Volumetric conjoint studies require respondents not just to indicate their product preference, but to specify how many units they would purchase. This introduces unique challenges around data quality, response realism, and modelling methodology.

Previous research (e.g., Eagle 2010[1]; Eagle, Louviere, and Islam 2018[2]; Hardt 2022[3]) has proposed a range of analytical methods. However, there remains no consensus within the industry, and barriers such as technical complexity, software limitations, and time constraints continue to hinder adoption.

This study replicates and extends these foundational ideas, testing five models:

- Naïve
- MNL MEV (Maximum Expected Value)
- Joint Discrete/Continuous (JDC)
- Polytomous Logit (via MBC software)
- HB-Reg (Hierarchical Bayes Regression)

The Volumetric Challenge

At first glance, volumetric conjoint may appear to be a straightforward extension of standard discrete choice methods, which involves simply asking people what they would buy and how many. However, analysts with experience in this area know that such studies introduce a host of design and data challenges that complicate both fieldwork and analysis.

Unlike binary or single choice questions, volumetric tasks require respondents to think more deeply about quantity decisions—something most consumers are not used to doing in a survey context. The concept of "how many would you buy?" seems simple, but without a clear shopping frame of reference, e.g., for what occasion, over what time period, at what budget, the question quickly becomes ambiguous. This cognitive ambiguity is one of the primary reasons volumetric studies often return inconsistent or unrealistic results.

These issues are further compounded by the design of the conjoint task itself. Asking respondents to report quantities across multiple SKUs, without necessarily having them consider substitution or total volume constraints, invites error. Some respondents overclaim, entering unrealistically high figures, e.g., 33 units, possibly caused by typos or interface confusion. Others "flatline" across tasks—defaulting to a uniform value, often one unit per product, simply to complete the exercise quickly. Then there are those who vary their inputs erratically between tasks, despite similar stimulus sets, reflecting either fatigue or lack of engagement.

Additionally, input mechanisms and survey platforms contribute to the challenge. Volume entry is often done through small number fields, drop-down menus, or manual text entry. On mobile devices especially, these mechanisms can frustrate users and reduce input accuracy. Positional bias, where SKUs presented near the top of the screen receive higher volumes, is also potentially more pronounced on mobile interfaces.

Critically, respondents often lack a reference point for how much volume is "normal" in the context being asked. Are they shopping for themselves or their household? For a weekend or a month? Is price a constraint, or are they being asked to choose freely? Unless these conditions are clearly communicated and reinforced throughout the exercise, the data returned can be highly variable and difficult to interpret.

These factors make volumetric conjoint a uniquely demanding methodology—not just in terms of modelling complexity, but also in respondent experience. Success requires careful design, rigorous data cleaning, and an understanding that asking for volume introduces significantly more variability and potential for error than simply asking for choice.

EXPERIMENTAL DESIGN

The case study was intentionally designed with a moderately complex design, in order to reflect the kinds of volumetric studies practitioners regularly face.

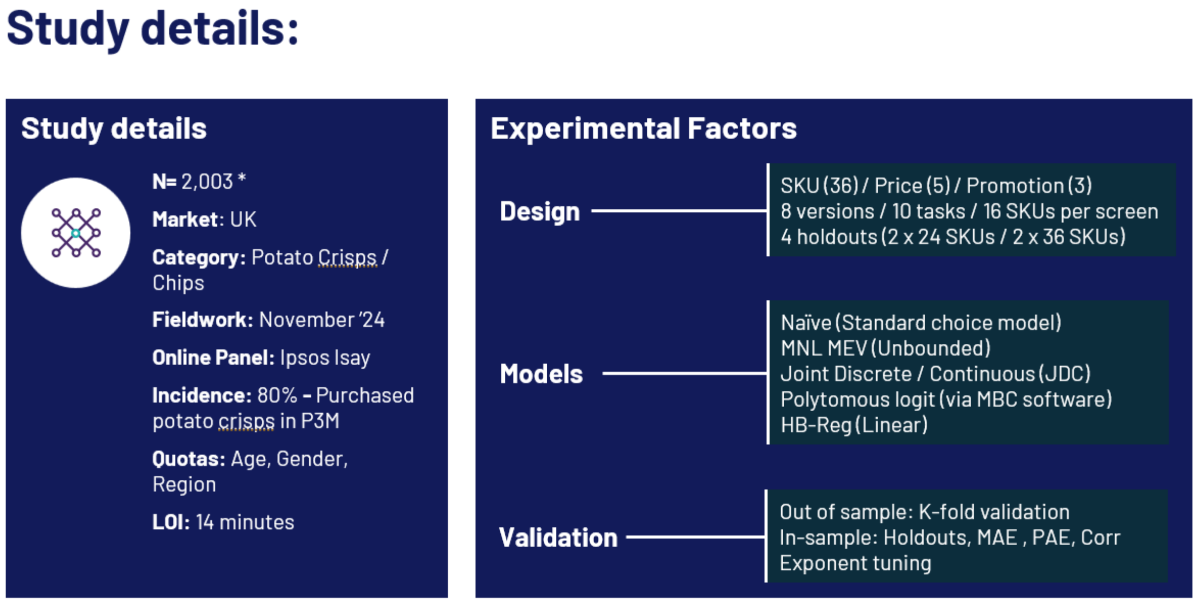
The focal category was potato crisps (or chips), and the fieldwork was conducted in the UK in November 2024. Participants were sourced through the Ipsos iSay online panel and respondents were eligible if they had purchased potato crisps in the past three months. Quotas were applied to ensure a nationally representative spread of age, gender, and region and after rigorous data cleaning, the final sample included 2,003 respondents.

The experimental design (Figure 1) consisted of 36 SKUs drawn from real market share data and each SKU had an alternative-specific price attribute (five levels per SKU). A promotion attribute was included that applied a generic percentage discount (0%, 25%, or 50%), with the no-discount level deliberately oversampled to appear in about 75% of cases.

Eight versions of the design were generated, and respondents completed 10 conjoint tasks each, with each task displaying 16 of the 36 SKUs. In addition, each respondent answered two holdout tasks. Half of the sample received two holdout tasks with 24 SKUs per task, while the other half saw 2 holdout tasks with 36 SKUs each, allowing for a meaningful test of how well the models extrapolate beyond the design. Using as few as eight versions would normally not be advocated, but to facilitate the out-of-sample K-fold validation work that was conducted, the number of versions was kept deliberately low.

The platform used was Sawtooth Software's Lighthouse Studio. The survey was designed to be device-agnostic, but all participants would have needed some level of scrolling regardless of the device used. Mobile respondents used a vertical scrolling interface. Respondents were instructed to enter the number of units they would purchase for each SKU based on the context of a typical "big shop" at a supermarket or online store. A None option was not explicitly programmed, but respondents could leave the task blank to indicate no purchase.

Figure 1



Five volumetric modelling techniques were evaluated in this study. These ranged from simple adaptations of traditional CBC models to more intricate, purpose-built volumetric solutions.

1. Naïve Model

The Naïve model applies a standard allocation choice model to the data. Each respondent's volume answers are converted to shares (totalling 100) and then preference shares are estimated using the logit model where the None option was coded in the data as a 1 if a respondent did not select any SKUs. To incorporate volume, the model multiplies these preference shares by the average number of units a respondent claimed to buy across all 10 random tasks. While basic, this approach benefits from simplicity and minimal setup.

2. MNL MEV

This method introduces the concept of Maximum Expected Value (MEV)—defined as the highest number of units a respondent selected across any task. The assumption is that this MEV represents the respondent's theoretical upper bound for purchase volume. In each task, the difference between the MEV and the total selected units is assigned to a synthetic "None" option and a logit model is run. This provides a flexible, unbounded interpretation of volume but may suffer if the MEV is skewed by unrealistic outlier tasks. Previous literature[6] has suggested artificially increasing the MEV value as respondents are highly unlikely to have evaluated a highly desirable combination of SKUs, prices and promotions. However, this was not done for this study. To incorporate volume, the approach is the same as in the Naïve model where the preference shares are multiplied by the MEV.

3. Joint Discrete/Continuous (JDC)

Proposed by Tom Eagle (2010)[1], the JDC method separates choice and volume into a two-stage model. The first stage estimates a logit model (as in the Naïve model). In the second stage, a regression model is used to predict the total number of units selected, using the natural log of the summed exponentiated SKU utilities as the independent variable. Taking the log helps, though not completely eradicates, the issue of there being a difference in the number of SKUs tested in a choice task versus the number of SKUs tested in a simulator, which tends to always be higher. The underlying assumption in this approach is that the total volume purchased in any task is directly related to the scenario's inherent appeal.

4. Polytomous Logit (via MBC)

This model was implemented using the Sawtooth Menu-Based Choice (MBC) software. The motivation for using this method is that in a standard logit model, even using HB, when the price of an SKU is changed, it gains or loses share from every other SKU in the simulation, which is not the market reality. While there are approaches that can reduce that effect, such as nested logit models, they are complicated. However, using the in-built functionality in MBC it is possible to identify partial effects, or in other words, identify a subset of SKUs that have the biggest impact on volume change and include only those SKUs in the modelling. An issue from a mathematical point of view is that a polytomous logit model treats volume as a categorical outcome. To work, it requires that each volume (e.g., 0,1,2,3,4,…) is its own category and a separate set of parameters is potentially estimated for each category. The advantage is flexibility: the model can estimate partial cross-elasticities and SKU-specific dynamics. However, it is mathematically imprecise, treating a continuous concept as categorical and is resource intensive. Each SKU requires its own model, and any cross-effects to be modelled have to be separately identified for each SKU.

5. HB-Reg (Hierarchical Bayes Regression)

This approach involves running a continuous regression model for each SKU, where the dependent variable is volume. In theory, this method should outperform MBC because it treats volume as a true continuous variable. For this study the volume was not transformed (e.g., by taking the logarithm), and the raw volume value was used. The same cross-effects identified from the MBC model were used for this model.

Data Quality and Cleaning

Ensuring data quality is critical for this type of study, particularly because volumetric conjoint tasks inherently introduce more variability and higher risk of poor-quality responses than traditional choice-based tasks. Respondents are required not only to engage with multiple product options per task, but also to specify quantities, adding additional cognitive load and increasing the chance of poor-quality data.

To manage these risks, a series of validation steps were applied. These included both real-time logic checks during the survey and post-hoc cleaning rules.

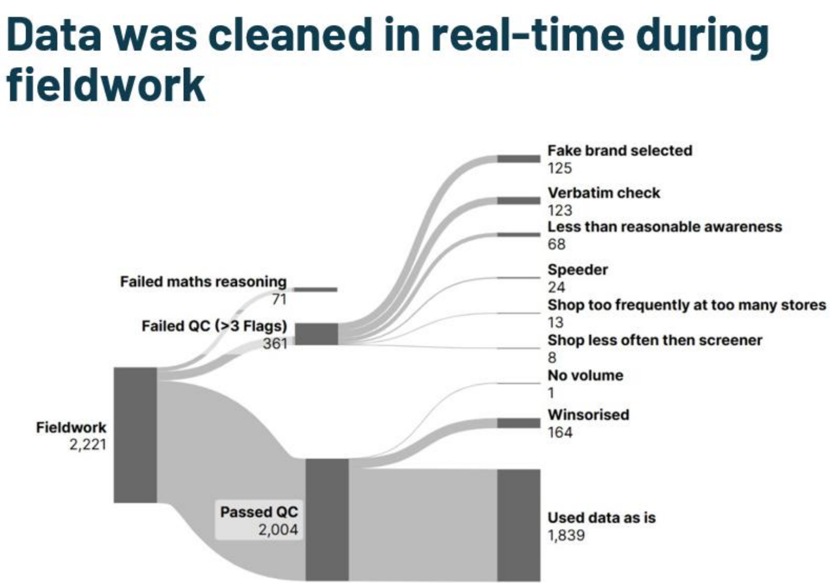
Within the questionnaire a math-based logic check question (Eden, Barkley, Olsen, 2024[4]) was used to identify potentially fraudulent respondents, or real respondents who might not be reading questions attentively.

Several flags were placed throughout the survey and if a respondent failed multiple flags they were removed post-survey. Flags included adding a fake brand to the brand list, a check on the awareness of the number of brands in the category, shopping frequency and when respondents last purchased potato crisps.

Open-end responses were included after the conjoint experiment and evaluated for nonsensical or bot-like answers. Speeding was another key indicator. Respondents were flagged if they completed the survey in under 40% of the median length of interview (LOI). While not always indicative of bad data on its own, when paired with other flags it often signalled low engagement.

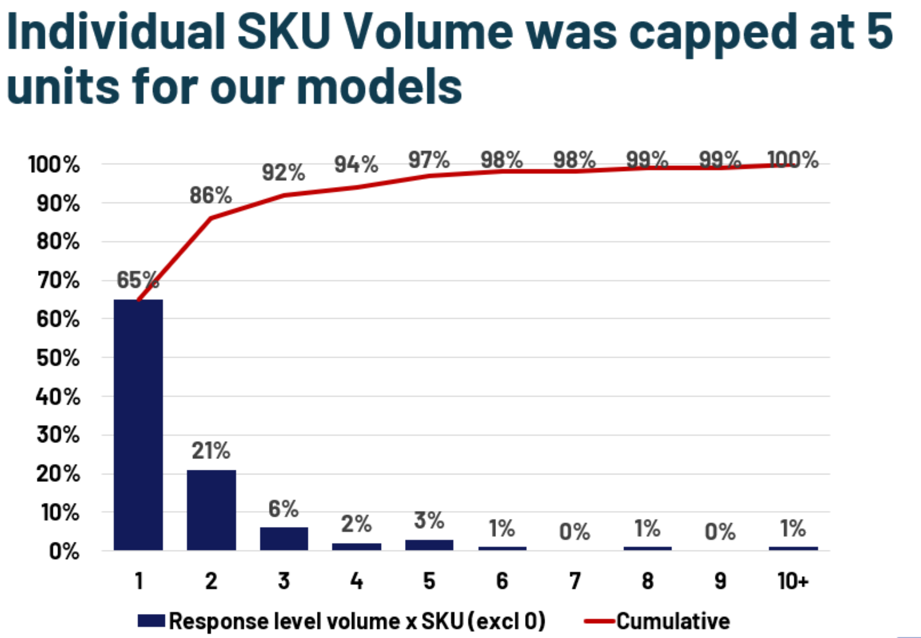
As shown in Figure 2, approximately 10% of the collected data was removed due to failing these quality checks. This substantial proportion underscores the importance of rigorous validation when dealing with volumetric tasks.

Figure 2



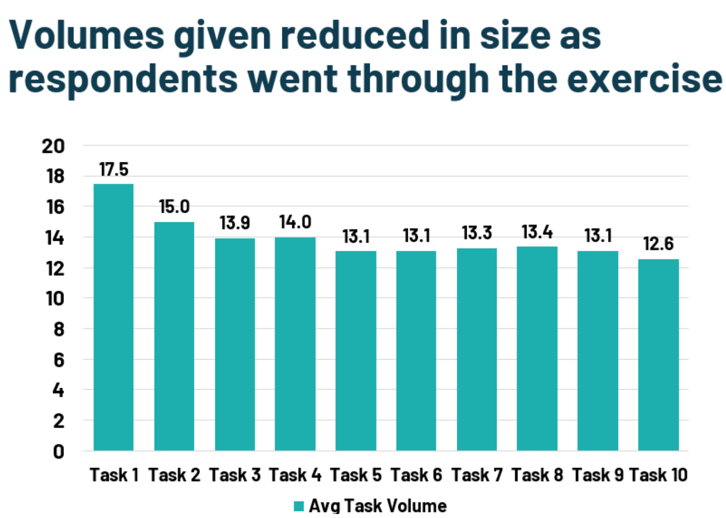
Another consideration is how to cap volume inputs. Based on the data (Figure 3), SKU-level volume was capped at 5 units as 97% of all SKU-level entries fell within this limit. This approach ensured the vast majority of the data was retained while limiting the influence of outliers. Task-level volume, i.e., the total volume selected across all SKUs in a single task, was not capped, although in 6% of cases the total task volume surpassed 30 so there may have been justification for capping. This, however, requires the analyst to manually adjust the SKU-level volumes within those tasks.

Figure 3



Respondents appeared to require time to settle into the experiment. When looking at the average volume by task it clearly shows higher volumes in the first couple of tasks before respondents settled into a more consistent range (Figure 4).

Figure 4

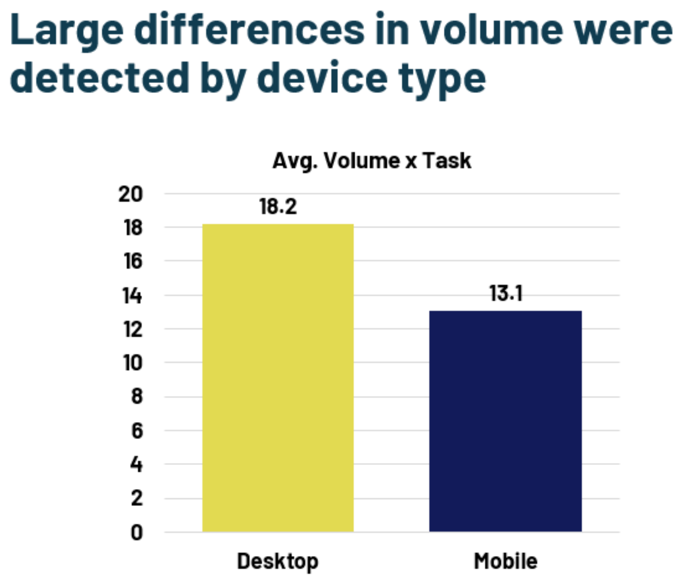


Previous literature[5] has suggested that respondents in general differ in their behaviour and response patterns in the early tasks, not just for SKU-based conjoint but for conjoint in general, and there are arguments for throwing away the first task or having a practice task before the main conjoint experiment.

Device type also played a major role in shaping data quality. In total, 68% of respondents completed the survey on a mobile phone, highlighting the need for device agnostic surveys. Some systematic patterns were observed where mobile users consistently entered lower volumes than those on a desktop/laptop, with mobile respondents averaging 13.1 versus desktop/laptop respondents who averaged 18.2 (Figure 5).

Again, previous literature[7] has shown marginal bias where mobile respondents are more likely to select the first option than desktop/laptop respondents. An analysis of the position data in this study showed that positional bias was less of a factor with all respondents over-indexing on the early SKU positions. While mobile respondents did over-index more on the first 8 positions there was little difference between the device types.

Figure 5



Analytical Setup and Challenges

Each model posed its own unique implementation and practical challenges, many of which reflect the real-world issues that practitioners face when conducting volumetric analysis.

For the Naïve and MEV approaches (Figure 6) the model was set up where a part-worth parameter is calculated for each SKU. Price was specified as an alternative-specific linear parameter for each SKU while the promotion attribute was coded as a single part-worth parameter. In the response column the volume is standardised to sum to 100 within each task. More complex analysis such as price tiering, where SKUs are grouped together, and a single price parameter is calculated for the tier was not conducted. Figure 6 shows a snippet of the data file setup.

Figure 6

sys_RespNum	Task	Concept	SKU	Price1	Price2	Price3	Price36	Promotion	Response
2	1	1	26	0	0	0	0	1	0
2	1	2	3	0	0	1	0	1	0
2	1	3	27	0	0	0	0	1	0
2	1	4	29	0	0	0	0	1	0
2	1	5	7	0	0	0	0	1	16.66667
2	1	6	35	0	0	0	0	1	0
2	1	7	24	0	0	0	0	1	0
2	1	8	11	0	0	0	0	1	0
2	1	9	9	0	0	0	0	2	0
2	1	10	30	0	0	0	0	1	0
2	1	11	14	0	0	0	0	1	0
2	1	12	6	0	0	0	0	1	0
2	1	13	8	0	0	0	0	3	0
2	1	14	1	3	0	0	0	1	0
2	1	15	18	0	0	0	0	1	0
2	1	16	13	0	0	0	0	1	16.66667
2	1	17	0	0	0	0	0	0	66.66667

The Joint Discrete/Continuous (JDC) approach requires a two-stage estimation process. The first model was set up exactly as the Naïve and MEV approach, then a second-stage regression model was conducted. From the first model the utilities for each SKU for each task are calculated to create the denominator, which is the natural log of the sum of exponentiated utilities for all SKUs in the task, and the outcome variable is the volume. An HB-Regression model was used for the analysis rather than a standard linear model as HB-Reg helps smooth out outliers via the HB shrinkage effect.

The polytomous logit model, implemented through Sawtooth's MBC platform, was very labour-intensive. For this model, a total of 36 separate polytomous logit models were required, one per SKU. It meant that availability effects need to be added to the model specification to understand the effect of adding or removing a SKU in the simulation tool. When an SKU is not included in a task, the price and promotion levels for that SKU are set to zero (Figure 7). As the software is expecting a categorical outcome variable it is not possible to specify a value of zero to represent when volume = 0, so the volume data is coded from 1 to 6.

Figure 7

Serial	SKU1a	SKU1b	SKU1c	SKU2a	SKU2b	SKU2c	SKU36a	SKU36b	SKU36c	DV1	DV2	...
2	1	3	1	2	0	0	2	0	0	1	1	...
2	2	0	0	2	0	0	1	1	1	1	1	...
2	2	0	0	2	0	0	2	0	0	1	1	...
2	1	3	1	1	2	1	1	2	1	1	2	...
2	1	5	1	2	0	0	2	0	0	1	1	...
2	2	0	0	1	5	2	2	0	0	1	1	...
2	1	4	1	1	3	1	2	0	0	1	3	...
2	2	0	0	2	0	0	1	4	1	1	1	...
2	1	1	1	2	0	0	2	0	0	1	1	...
2	2	0	0	2	0	0	1	5	2	1	1	...
3	1	2	1	1	4	1	2	0	0	2	1	...
3	2	0	0	2	0	0	1	3	1	1	1	...
3	1	5	1	1	1	1	2	0	0	1	1	...
3	2	0	0	1	1	1	1	5	1	1	1	...
3	2	0	0	1	4	1	2	0	0	1	1	...
3	2	0	0	2	0	0	2	0	0	1	1	...
3	1	4	1	2	0	0	1	3	1	1	1	...
3	1	1	1	1	5	1	1	2	2	1	1	...
3	2	0	0	2	0	0	2	0	0	1	1	...
3	1	2	2	1	3	1	2	0	0	1	1	...

SKU(X)a = Availability effect
SKU(X)b = Price effect
SKU(X)c = Promotion effect

To create a fully specified model, 535 parameters would need to be estimated for each of the 36 SKU models. As previously indicated, a key reason for looking at this model is the ability to specify a partial effects model and only include those parameters, or SKUs, that matter and have an effect on volume. To identify the partial effects, an aggregate logit model was run for each SKU model where all 535 parameters were specified. From the resulting output, parameters were selected based on not only the size of the effect and the parameters being of the correct sign/direction but also expert category knowledge (e.g., is the competing SKU part of the same brand, or premium-ness, or in the same price backet). After completing that exercise for each SKU, most of the individual SKU models contained approximately 40 to 45 parameters. The parameters specified were different for each SKU model because each SKU has a different competitor set.

The HB-Regression approach was the most challenging model of all. As for the polytomous logit model, this technique required running separate HB models for each SKU, with volume as the dependent variable. However, the HB-Reg software lacks some of the MBC software functionality which caused significant issues. Within the MBC software there is functionality that allows conditional dependencies. This means the data is only submitted once, then in the software the user specifies that for each SKU model to only include tasks where that SKU is available. As HB-Reg doesn't have this functionality, a separate data file must be created for each model.

HB-Reg, unlike MBC does not allow batch processing, meaning each of the 36 models had to be configured and run individually. The software also struggled when certain SKUs lacked variation in attributes, such as promotion level requiring that some parameters had to be excluded. HB-Reg also does not output a simulator meaning additional work for the analyst to run the results.

These constraints speak to a broader truth in volumetric conjoint: while sophisticated methods are available, they often demand more time and technical skill than analysts can feasibly allocate.

TERMINOLOGY

The performance of each of the models was analyzed through three key lenses.

Absolute (ABS) Volume Difference—this metric states how far off the predictions for volume are, at a total task level. A large value here indicates that the model severely under or over predicted total volumes. A value of 1.5 for this measure would indicate predictions were 1.5 units outside that of the holdout data.

Volume Percent Average Error—is an indication of how well the model predicts SKU level volumes, and on average how far off as a % the model is. Lower percentages here indicate that the model performs better at predicting individual SKU level volumes.

Share Percent Average Error—is similar but instead looks at the predictions of SKU level volume share, and on average how far off as a % the model is. Lower percentages in this metric indicate if the share of SKU volume is close, even if the actual volumes themselves are not.

These PAE measures are calculated by taking the absolute volume/share difference at a SKU level and dividing it by the average level of volume/share across the entire dataset. For example, if on average a SKU has a 10% share of preference, and the model predictions are on average 2.5% off, then the Share PAE would be 25%.

Results

The first test involved undertaking out-of-sample K-fold validation using each of the 8 design versions. Here, each model's ability to predict task responses was assessed by comparing predicted vs. observed outcomes across all ten tasks in the version.

The Naïve model, despite its simplicity, performed well on key volume metrics (Figure 8). It was particularly strong when predicting share of volume and ranked third when it came to absolute volume difference. JDC and HB-Reg outperformed others in total volume prediction, indicating that the regression-based volume estimation approaches captured broader patterns of respondent behaviour effectively. The MEV and MBC models, while not the best in total volume prediction, did better at allocating share across competing SKUs.

Figure 8

	ABS Volume Diff	Volume PAE	Share PAE
NAÏVE	0.75	24%	19%
MEV	1.08	25%	15%
JDC	0.70	30%	27%
MBC	1.67	34%	20%
HB-REG	0.71	30%	25%

When respondents were exposed to larger SKU sets in the in-sample 24 SKUs per task holdouts, the results changed (Figure 9). JDC remained strong, particularly for volume metrics, while MBC offered reliable performance across both share and volume. HB-Reg, however, began to show instability. As the number of products in the scenario increased, HB-Reg's performance dropped sharply, due to the model's failure to adapt when simulations included more SKUs than the model was estimated on.

Figure 9

	ABS Volume Diff	Volume PAE	Share PAE
NAÏVE	0.57	18%	16%
MEV	0.76	16%	14%
JDC	0.21	17%	17%
MBC	0.42	23%	19%
HB-REG	2.55	51%	22%

The most challenging scenario involved the 36 SKUs in-sample holdouts, a situation which is likely to be closer to how real-world simulations compare to the tasks the models are trained on. MBC emerged as the top performer, preserving both share and volume accuracy. Naïve and HB-Reg models struggled significantly, failing to manage the increase in volumes which came about due to the larger number of SKUs in the holdouts. The JDC model continued to perform well, ranking second in its ability to predict absolute volume and SKU volume (Figure 10).

Figure 10

	ABS Volume Diff	Volume PAE	Share PAE
NAÏVE	2.26	30%	11%
MEV	1.79	24%	11%
JDC	1.26	22%	18%
MBC	0.64	19%	17%
HB-REG	2.79	53%	29%

The analysis of "volumetric drift," a measure of how well each model performed as the number of SKUs increased, showed that the Naïve, MEV, and HB-Reg models showed greater error as SKU count rose, suggesting that their assumptions break down under scale. MBC and JDC predictions, however, remained fairly stable, with MBC improving due to the increased richness of SKU cross-effects.

The final experiment involved undertaking "exponential tuning"; a technique used to adjust utility scaling before exponentiation in simulations. From the tests conducted it was found that reducing the exponent to between 0.60 to 0.75 yielded consistent improvements across most models, particularly for JDC and MBC (Figure 11). This relatively simple adjustment to the model had a large impact on reducing the error in our predictions yet remains largely unused in many conjoint applications.

Figure 11

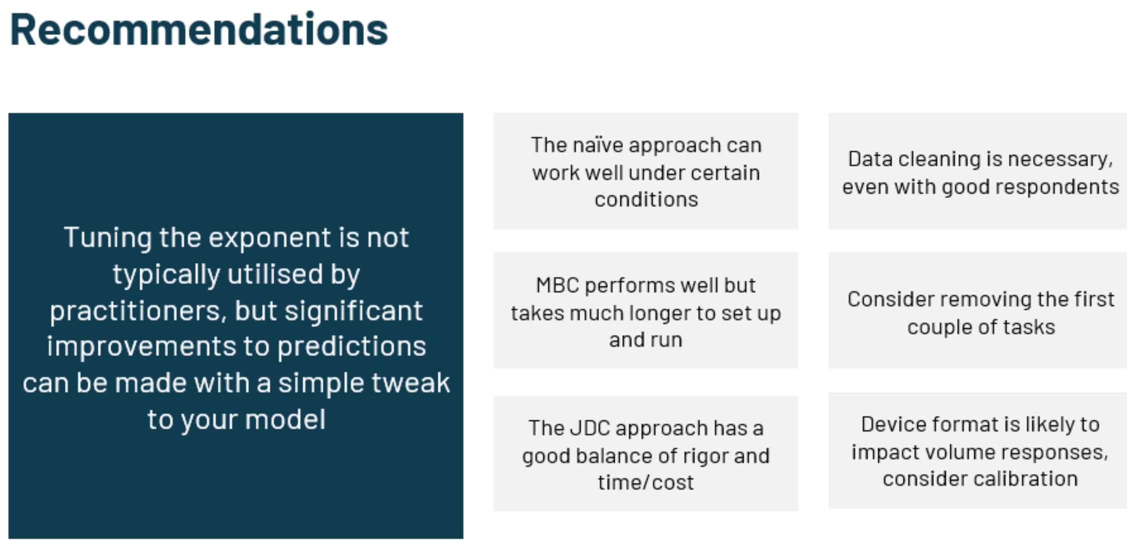
OSS K-Fold	1.0	0.75	0.60	0.50
NAÏVE	22%	21%	22%	24%
MEV	25%	24%	34%	43%
JDC	30%	23%	21%	21%
MBC	34%	29%	24%	25%
HB-REG	24%	48%	68%	81%

Recommendations for Analysts

Several key insights emerged from this study (Figure 12) that are directly applicable to day-to-day volumetric practice:

1. **Data Quality is Paramount:** Even respondents who appear engaged can provide erratic volume inputs. Use logic checks, remove outliers, and apply sensible caps to improve model reliability.
2. **Discard Early Tasks:** The first few tasks often contain inflated or unrealistic volumes. Consider removing these from the estimation file or introducing practice tasks before you begin your exercise.
3. **Understand Device Effects:** Mobile respondents behave differently—not just in how they interact with the interface, but in how they report volume. It may be appropriate to calibrate volumes across devices to get a consistent view on volume.
4. **Choose the Right Model for Your Needs**: The Naïve model is fast and easy to implement. Use it when volume is stable, and the number of SKUs you are simulating is the same as the number of SKUs used to train your model. MBC, whilst showing promise requires a lot of data wrangling and therefore may not be appropriate. JDC hits a sweet spot between practicality and power and would be the recommendation to most analysts.
5. **Calibrate with Exponents:** Whilst not often used by analysts, there was an impressive reduction in Volume PAE when tuning the models with a factor of 0.6. Tuning models is a relatively simple approach which can have potentially large impacts on accuracy.

Figure 12



Conclusion

Volumetric conjoint analysis is an underused technique in the practitioner's toolkit. Whilst traditional choice-based approaches help understand preferences, they fall short of providing the more granular demand estimates which some stakeholders demand. However, as this study shows, the potential of volumetric conjoint comes with considerable practical and methodological challenges.

This research has shown that while numerous modelling options exist, their effectiveness varies significantly depending on the complexity of the task, the size of the SKU universe, and the quality of the input data. Simpler models like the Naïve approach still hold surprising value, especially in well-designed studies with stable volumes. However, their limitations become apparent when trying to scale or simulate more dynamic real-world scenarios. On the opposite end, models like HB-Reg, though theoretically appealing can collapse.

MBC and JDC approaches emerged as the strongest overall contenders. JDC provides solid performance across both volume and share metrics, and once implemented, is scalable with relative ease. MBC, while showing promise across larger SKU sets requires a substantial setup phase.

This study also helps to highlight a broader truth of volumetric modelling: good data will always be more important than our modelling technique. The most powerful algorithm in the world cannot compensate for poorly engaged respondents or unchecked outliers. From task design to device usability to cleaning protocols, getting the data right is half the battle in volumetric studies.

Ultimately, turning it to eleven doesn't always mean using the most complex approach. It means choosing the right level of sophistication for your goals and balancing methodological ambition with practical realism.

Dean Tindall

Chris Moore

Manjula Bhudiya

References

1. Modelling Demand Using Simple Methods—Tom Eagle, Sawtooth Software Conference 2010 Proceedings p283 https://sawtoothsoftware.com/resources/technical-papers/conferences/sawtooth-software-conference-2010
2. A Comparison of Volumetric Models—Tom Eagle, Jordan Louviere & Towhidul Islam, Sawtooth Software Conference 2018 Proceedings p267 https://sawtoothsoftware.com/resources/technical-papers/conferences/sawtooth-software-conference-2018
3. Volumetric Conjoint and the role of Assortment Size—Nino Hardt, Sawtooth Software Conference 2022 Proceedings p251 https://sawtoothsoftware.com/resources/technical-papers/conferences/sawtooth-software-conference-2022
4. Yoshimi Battles the Survey Bots—Layla Eden, Daniel Barkley & Trevor Olsen, Analytics & Insights Summit 2024 Proceedings p255 https://sawtoothsoftware.com/resources/technical-papers/conferences/analytics-and-insights-summit-2024
5. How Many Questions Should You Ask in Choice-Based Conjoint Studies?—Richard M. Johnson and Bryan K. Orme, Sawtooth Software, Inc., 1996 ART Forum, Beaver Creek, https://content.sawtoothsoftware.com/assets/a24654f4-0553-4484-9c90-46a5899e8d57
6. Becoming an Expert in Conjoint Analysis 2nd Edition—Bryan Orme & Keith Chrzan 2021 p171
7. Choice Based Conjoint in a Mobile World—How far can we go?—Chris Moore and Christian Neuerberg, Sawtooth Software Conference 2016 Proceedings p97 https://sawtoothsoftware.com/resources/technical-papers/conferences/sawtooth-software-conference-2016

THE WILL OF THE MANY: GENERATING NOVEL CONCEPTS USING AI-ENHANCED RESPONDENT FEEDBACK

JORIS VAN GOOL
PETER LI
SKIM

1. INTRODUCTION

Successful new product design (NPD) requires a thorough understanding of the preferences of its intended target audience. For this, market researchers have adopted survey-based methods, such as conjoint analysis, as the standard for the in-depth evaluation of new concepts. To accurately capture consumer preferences, however, these methods require that the included designs encompass all potential designs that may appeal to respondents (Cunningham, et al., 2010). In practice, this often proves challenging.

Conjoint analysis starts with defining an attribute-by-level grid that describes the product or service design space. This approach proves highly effective when the attributes are known and well-defined, and when the number of levels remains compact. However, within the context of creative design, such as package design, the methodology encounters significant constraints. Namely, the space of potential designs becomes vast and high-dimensional, making it difficult to impose meaningful attribute-level grids from the outset. This a priori structuring risks excluding design elements that consumers might find particularly attractive, since the framework may not anticipate all relevant creative possibilities. Mitigating this risk often requires extensive preliminary research and provides no guarantee of effectiveness.

To address this, we first posit that we can attain better coverage of high-potential designs by providing respondents with full freedom regarding the composition of their ideal design. We achieve this by providing them with text-to-image models, such that they can instantaneously translate their ideas into designs. We postulate that while most respondents lack the requisite skills to visualize a design, they can accurately formulate their preferences through natural language. We subsequently test this hypothesis by comparing the quality of our AI-generated designs to each other and to professionally designed ones in a respondent-level MaxDiff study.

Beyond the generation of novel designs, we leverage the respondent prompts to identify the constituent elements of a successful design. To achieve this, we leverage a multi-modal language model to decompose each respondent-generated design into a set of human-interpretable attributes. These attributes then serve as the basis for estimating an attribute-level choice model, permitting us to quantify the relative importance of each attribute in driving respondent preference.

In our paper, we propose a novel methodology inspired by crowdsourcing, utilizing state-of-the-art AI models for creating superior product designs. We empirically validate our approach by conducting a large-scale survey with Christmas-themed Coca-Cola cans as the product of interest. We furthermore provide practical guidelines that describe the limitations and boundary conditions of our approach. The remainder of the paper is outlined as follows: We first discuss how our paper contributes to existing research; we then describe our methodology and data. We subsequently present and discuss our results. Finally, we summarize our findings in the conclusion and present fruitful avenues for future research.

2. Background

In literature, a consumer-driven approach for NPD, where consumers are asked to either propose new designs or suggest improvements to existing ones, is commonly referred to as crowdsourcing designs. Past research has thoroughly investigated the benefits and trade-offs of this approach. Among others, Poetz and Schreier (2012) show that user-suggested ideas score significantly higher concerning novelty and customer benefit, although at the expense of a degree of feasibility. Allen et al. (2018) reinforce these findings by showing that crowd-sourced designs were positively correlated with sales, contingent on the original design being of sufficient quality. Using a case study of crowd-sourced clothing designs Piller (2010) further emphasizes the necessity of a lower bound on the quality of the sourced designs. Although their research highlights a substantial improvement on numerous metrics, they stress the fact that many of the participants were semi-professional designers.

In the absence of this, we utilize AI-generated designs instead. The class of text-to-image models allows us to instantaneously translate natural language descriptions into images. This has only been made possible recently through great advances in computer vision and machine learning, such that state-of-the-art models are now able to generate images from human prompts with sufficiently high fidelity (Rombach, et al., 2022).

As image generation models have only recently achieved a satisfactory level for practical usage, applications within market research are still scarce. Nevertheless, Hartmann et al. (2024) conduct a large-scale study on AI-generated visual marketing content. In their paper, they compare their AI-generated designs against professionally designed ones. They show that their designs outperform professional designers in certain cases. While their paper addresses the need for expensive designers and the associated time-consuming process, they do not tackle the lack of exploration within the design space. Rather, they wholly rely on their judgment and therefore their inherent biases to conceive prompts.

Our paper contributes by simultaneously focusing on exploitation and exploration within the design space. First, we exploit high-potential areas through leveraging the inherent stochasticity of AI models to generate multiple unique designs for each prompt. We then thoroughly explore the space by asking a large and diverse group of respondents to describe their ideal design, allowing us to draw inspiration from across the entire design space.

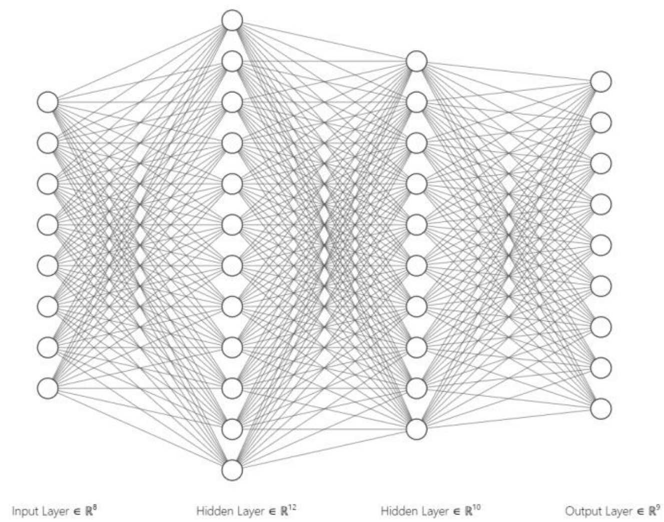
3. Image Generation

In this section, we elaborate upon the design generation element of our research. We first give a brief non-technical introduction to artificial intelligence (AI) and text-to-image models. We then elucidate our process for selecting the model to be used by respondents. Finally, we will explain the details of our field experiment.

3.1 Deep Learning

Artificial intelligence (AI) is an all-encompassing term that broadly defines any system or machine capable of mimicking human intelligence. While often used interchangeably with machine learning (ML), machine learning is a subset of AI. More specifically, ML is a form of AI that focuses on teaching systems to learn patterns from data using algorithms and computational models, rather than being explicitly programmed. Within the realm of machine learning resides deep learning. Deep learning generally refers to a class of machine learning models known as neural networks, which are designed to represent data through multiple hierarchical layers of abstraction. Many state-of-the-art machine learning models are now based on some neural architecture. This includes virtually all of what is now referred to as Generative AI (GenAI) and, therefore, image generation models as well.

Figure 1: Stylized representation of a four layer fully-connected neural network.



We depict a high-level representation of the most basic neural network architecture in Figure 1. The actual model we use for generating our designs is substantially more complex; however, the core concepts remain identical. Consequently, we use this simplified example to illustrate the process. Each node or neuron in the figure is equivalent to a real-numbered scalar. We see that our input and output layers are vectors of dimension R^8 and R^9 respectively. If we equate each neuron in the input layer to a single word, and each neuron in the output layer to a pixel of an image, we can then translate a sentence of up to 8 words into an image of up to 9 pixels. To go from input to output, we sequentially move through our hidden layers. Each neuron within a hidden layer is a linear combination of all neurons in the previous layer, including a scalar bias term with a non-linear transformation applied to it.

$$y_i = f(X\beta_i + b_i)$$

Here, y_i is the ith neuron in the current layer. X is the vector of all neurons in the previous layer. β_i are the weights corresponding to neuron y_i and b_i is the scalar bias term. $f(\cdot)$ can be any arbitrary function, although a commonly used one is SwiGLU (Shazeer, 2020). We initially start with a random initialization for the weights and biases. By providing labeled pairs of sentences and images, we can compute the loss, typically cross-entropy or mean squared error, between the generated images and the corresponding ground truth. By applying a numerical optimization method, we can minimize the loss across our entire training sample.

3.2 Model Selection

Model training is among the most important determinants of model performance. Since training competitive AI models has become prohibitively expensive, often involving tens of thousands of GPU hours, practitioners tend to depend on pre-trained models provided by top research labs. In this subsection, we explain how we find the most suitable among these models for our objective. For this, we first give a brief primer on how these models are pre-trained and then expand on how we as practitioners can further adapt and build upon them for our specific needs.

3.2.1 Unsupervised Pre-Training

State-of-the-art image generation is predominantly achieved by using a class of neural networks known as diffusion models, as introduced by Sohl-Dickstein et al. (2015) and popularized by Ho et al. (2020). While our toy example in Figure 1 included only 8 + 12 + 10 + 9 = 37 neurons and 337 weights and biases, in practice, these models often contain billions of parameters and are trained on text and image pairs scraped from the entire public internet. This process of imparting vast amounts of information into the model without explicit supervision is called pre-training. This paradigm of training increasingly large models on increasingly large data sets was popularized by Radford et al. (2018) and empirically proven beneficial by Kaplan et al. (2020), albeit for language models.

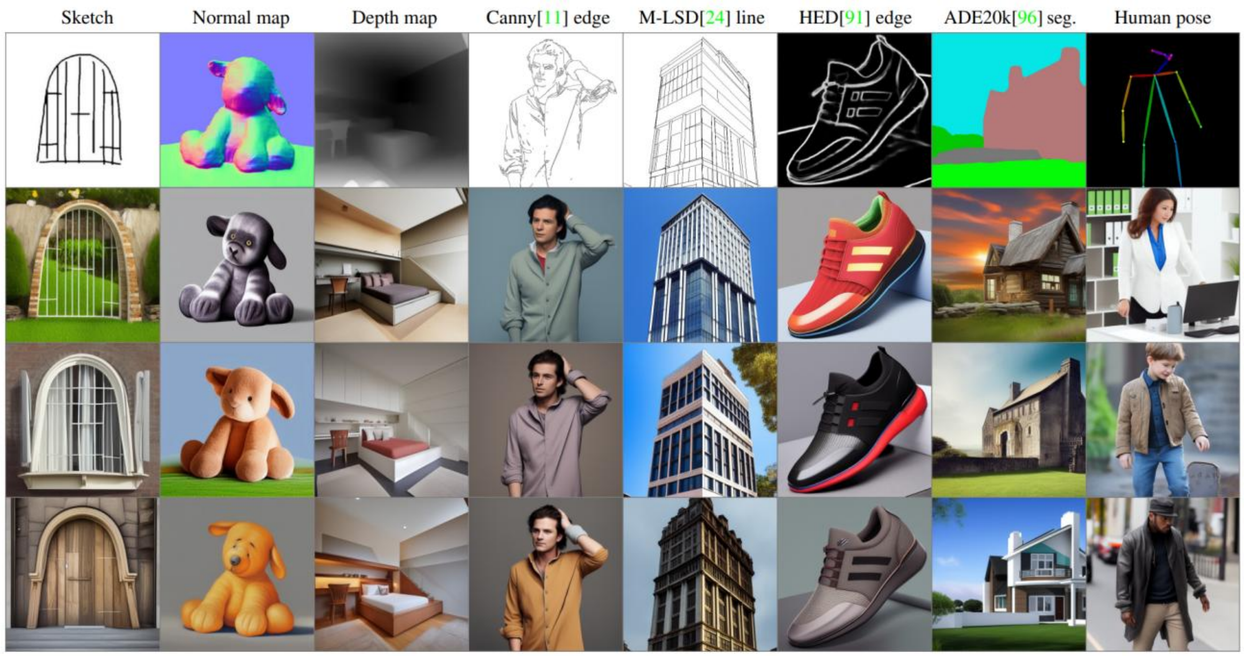
Pre-training has shown remarkable effectiveness in practice, resulting in general-purpose models capable of generating high-fidelity images for a wide array of prompts. The performance of these models, however, generally tends to collapse at tasks that were not included in the pre-training. Virtually all state-of-the-art models are either closed source or open weights, meaning that their training data is not publicly accessible. Consequently, it is impossible to determine a priori whether our exercise is included in the pre-training and, therefore, whether the model can perform it adequately.

3.2.2 Fine-Tuning

The aforementioned limitation of pre-training highlights the need for further adaptation on downstream tasks. While pre-training aims to imbue models with broad world knowledge, often by scraping vast amounts of internet data that may not be perfectly formatted or directly relevant to specific tasks, it cannot cover every conceivable application. To remedy this, we can employ further downstream training on a smaller, high-quality set of data. The intuitive interpretation here is that the model uses its imparted worldly knowledge to serve as a starting point for specialization.

3.2.3 ControlNet

Figure 2: Examples of ControlNet from Zhang et al. (2023).
By providing a template, we can fix the composition of the generated image.



However, even with fine-tuning, achieving precise control over specific compositional requirements within the generated image, such as brand-specific design features for products, can remain challenging. While models can acquire these capabilities through extensive fine-tuning, this typically demands a large, high-quality dataset, which is often unavailable. To resolve this issue, Zhang et al. (2023) propose ControlNets to allow image models to accept additional requirements that the generated image must satisfy. ControlNets do not refer to any specific architecture, but rather to an approach for adapting any pre-trained model. Figure 2 illustrates a few examples of how they function in practice. For the exact technical details, we refer the reader to the respective paper.

3.3 Experiments

To determine which approach works best for our problem, which is the design of a Christmas-themed Coca-Cola can, we conduct experiments with all three of the aforementioned methods. We manually evaluate the generated designs using a group of internal respondents.

3.3.1 Setup

Table 1 lists the configurations we tested. Fine-tuning and ControlNets are only possible with open-weights models or when the creators included the adaptation themselves, hence the subset. We selected these models since, at the time of conducting these experiments, they were the top-ranked models based on the image generation leaderboard[1].

Table 1: Tested Configurations

Model	General-purpose	Fine-Tuning	ControlNet
Imagen 3	x		
Recraft V3	x		
Flux-Dev	x	x	x
Flux-Pro	x		x
Stable Diffusion 3.5-Large	x	x	
Stable Diffusion 3-Medium	x	x	x

For fine-tuning, we collect a set of publicly available custom Coca-Cola designs, which we manually label. For ControlNet, we create a Canny edge map from the standard Coca-Cola can and provide it as the additional input.

[1] Text To Image Leaderboard - a Hugging Face Space by Artificial Analysis

Figure 3: Left, sample of a Christmas-themed Coca-Cola can included in the fine-tuning sample.
Right, the Canny edge map we used to fix the composition of our generated image.

3.3.2 Fine-Tuning Results

Figure 4: Output from before fine-tuning (left) and after (right).

Figure 4 presents an example of an image produced from a fine-tuned Flux Dev. Notably, it bears a striking resemblance to the image from the training data. This appeared to be the most common pattern we observed. Namely, the fine-tuned model would limit itself to stylistic elements from the training set.

We inferred from this behavior that achieving comprehensive coverage of the desired design space would necessitate a substantially larger fine-tuning sample. Given the practical infeasibility of acquiring potentially thousands of professionally created designs, we concluded further experimentation with fine-tuning to be impractical for our objectives.

3.3.3 ControlNet Results

Similarly, Figure 5 illustrates an image produced by Flux Canny-Pro, with the same prompt, using the Canny edge map derived from a standard Coca-Cola can as an additional input. Our manual evaluation revealed that this sample is exemplary for the issues we encountered with ControlNet type models. Namely, although they achieved strong compositional adherence, they suffered from low fidelity and otherwise poor prompt adherence.

Figure 5: A sample output produced by Flux + ControlNet.

3.3.4 Base Model Results

As discussed in Subsection 3.1, AI models are trained on pairs of labeled data. In our case, this process uses pairs of images and their corresponding textual description to teach the model to effectively map a textual description to its associated image. However, a product of this training is that these models develop biases towards specific prompt formats. The task of identifying the most effective prompt format for a given objective is known as prompt engineering. Researchers have dedicated considerable efforts towards devising optimal prompt engineering strategies. Nonetheless, it is still considered more of an art than a science, requiring much manual trial and error with each model possessing its idiosyncrasies.

Through extensive prompting and manual evaluation across all models, we determined that Recraft V3 consistently delivered the best performance, provided its specific prompt format was adhered to accordingly. We present a comparison of images generated by Recraft and Flux in Figure 6.

Figure 6: Two images sampled from Recraft V3 (Above) and Flux-Pro (Below) respectively for the prompt: "A can of Coca-Cola.
The design of the can features Santa under a beautiful starry night.
The can sits in front of a white background." In contrast to Recraft, Flux struggles much more with generating a consistent white background and the Coca-Cola motif.

3.3.5 Prompt Rewriting

While Recraft V3 demonstrated strong performance contingent on precise prompt adherence, in practice, we determine it unrealistic that respondents would be able to consistently comply with such specific prompt formats, even when given detailed instructions. To address this, we rewrite all respondent prompts using a large language model (LLM). Specifically, we use GPT-4o-mini to minimize response times to prevent undue delays for respondents. The LLM, however, in turn also requires instructions to rewrite the respondents' prompts. Drawing upon our experience from prompt engineering, we test the following four prompts for GPT-4o-mini.

1. You will be given a list of ideas which should be incorporated into the design of a can of **Soda**. You are to edit the prompt such that it is grammatically correct. Furthermore, the rewritten prompt must start with "A can of **Soda**. The design of the can features:" and end with "The can sits in front of a white background."
2. You will be given a list of ideas which should be incorporated into the design of a can of **Coca-Cola**. You are to edit the prompt such that it is grammatically correct. Furthermore, the rewritten prompt must start with "A can of **Coca-Cola**. The design of the can features:" and end with "The can sits in front of a white background."
3. You are a creative designer. You will be given a list of ideas which should be incorporated into the design of a can of **Soda**. You are to summarize these ideas into a concise description. It is of the utmost importance that the description remains succinct and unambiguous, leaving no room for interpretation. The rewritten prompt must start with "A can of **Soda**. The design of the can features:" and end with "The can sits in front of a white background."
4. You are a creative designer. You will be given a list of ideas which should be incorporated into the design of a can of **Coca-Cola**. You are to summarize these ideas into a concise description. It is of the utmost importance that the description remains succinct and unambiguous, leaving no room for interpretation. The rewritten prompt must start with "A can **of Coca-Cola**. The design of the can features:" and end with "The can sits in front of a white background."

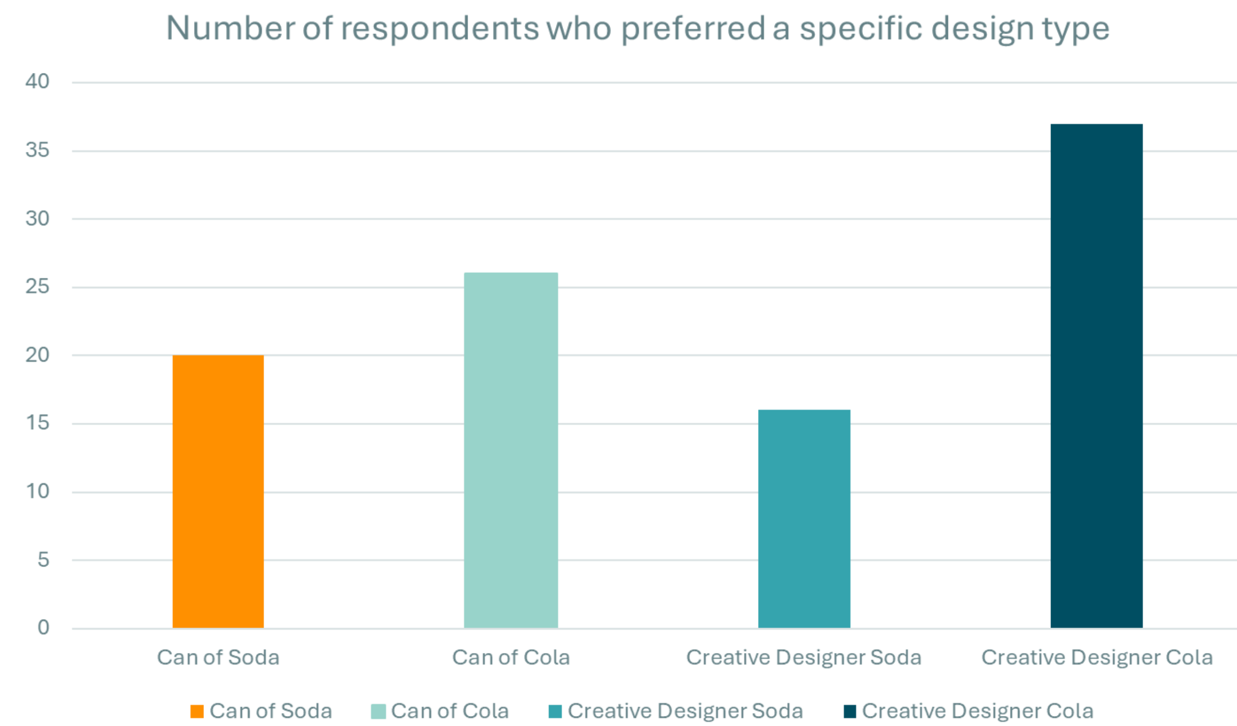
Importantly, we first distinguish between a generic can of soda and a can of Coca-Cola. We do so because we found in our initial experiments that the model occasionally struggled with consistently generating the Coca-Cola motif, two examples of which are shown in Figure 7.

Figure 7: Examples of generated cans where the Coca-Cola motif is poorly integrated.
For example, in the above examples it is either (partially) off-screen (left) or misshapen (right).

Furthermore, in our internal survey, we found that respondents primarily enumerated attributes that they want to appear on the can. We hypothesize that these text-to-image models would perform better when prompted with coherent sentences. In essence, we aim to leverage the LLM to partially automate this prompt construction process.

To ascertain the best rewrite prompt to be used in the real survey, we conduct a smaller survey with n=91. For each prompt, we show the user four designs, one for each prompt. We then ask the respondent to select their preferred design. Figure 8 shows us that respondents prefer it when the LLM is creative and inherently integrates the Coca-Cola motif onto the can. Accordingly, we proceed with prompt 4 exclusively in the real survey.

Figure 8: Distribution of respondents who preferred a specific design type.



3.4 Survey

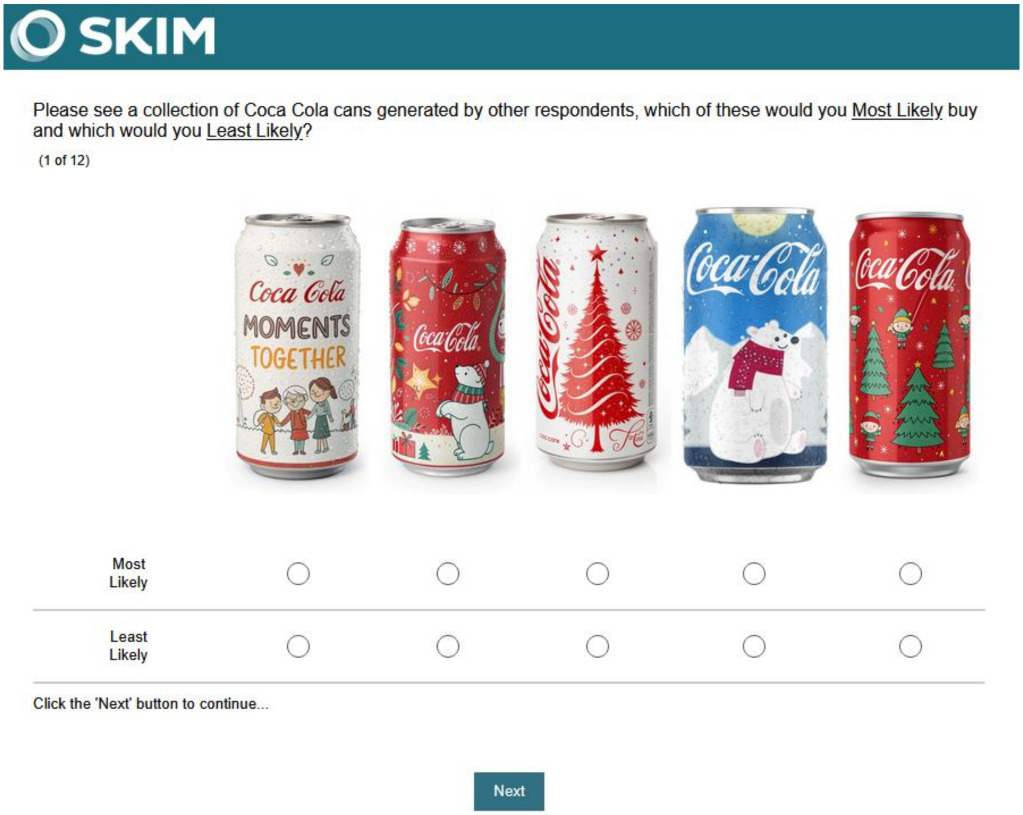
We carry out a survey with n=1300. All respondents are adults from the US who have consumed or purchased Coca-Cola within the past three months. We provide all respondents with the opportunity to submit up to three prompts. For each prompt, we generate four images. We do this to mitigate the effects of the inherent stochasticity of AI models. This approach increases the likelihood that respondents will find at least one image satisfactory. The selected images are then presented to other respondents in a subsequent MaxDiff exercise.

In the MaxDiff section, respondents are shown 25 designs generated by other respondents. Additionally, they are shown a sample of 5 out of 15 professional designs. These 15 designs were created by freelance designers whom we contracted. At each task, respondents are shown 5 out of the 30 available designs and requested to select their least and most favorite design, for a total of 12 tasks.

At the end of the MaxDiff section, we included an anchoring exercise where respondents were asked to indicate their willingness to buy cans they had seen in the MaxDiff over the regular Cola can. We used on-the-fly calculations to show the 1st, 7th, 13th, 19th, 25th, and 30th designs. This provides us with an anchor and increases the validity of the model fit KPI.

To ensure that no inappropriate designs are shown to respondents through inclusion in the MaxDiff, we remove the designs of respondents who prompt for violent, offensive, or otherwise unsafe elements. We do this through LLMs that we instruct to inspect for any of these elements. We verify both the prompt and the generated image.

Figure 9: The MaxDiff screen in our survey.
Respondents are asked to select their most and least favourite design out of the five shown.



4. Modeling

To translate the individual preferences obtained from our survey into aggregate preference rankings of Coca-Cola cans, we employ two distinct methodologies for modeling the choice data. The first approach is aimed at achieving the highest model fit, providing us with the most accurate aggregate ranking. The second endeavors to assist us in discerning the key elements that constitute a preferred design.

To attain the best fit, we estimate an anchored respondent-level choice model where we utilize a standard hierarchical Bayes model with a single modification. Namely, for model tractability, we implement a factored covariance matrix as in Lattery et al. (2025), adopted for our framework. This alleviates the considerable computational burden introduced by the large number of respondents.

In contrast, to understand which attributes drive consumer preference, we decompose the prompts into their constituent attributes and use them in a downstream attribute-level choice model. Conventionally, market researchers code attributes from open-ended surveys manually by hand. However, given our large sample size, this is a time-consuming and potentially error-prone process. To overcome these challenges, we propose utilizing large language models for this process. LLMs have achieved superhuman intelligence on numerous benchmarks (Gemini, et al., 2023). As a result, we posit that they should be capable of performing similarly to humans, if not better. To test this hypothesis, we perform the coding by hand as well and use this as a benchmark. Additionally, text-to-image models might not always include every element a respondent prompted for. Consequently, to account for potential discrepancies between respondent requests and what was actually displayed in the MaxDiff section, we additionally extract attributes from the generated images. For this, we utilize a multi-modal LLM, capable of interpreting both images and text.

We evaluate the relative performance of all our approaches by comparing their RLHs.

5. Results

In this section, we will dive into the results. We will show the design features identified from respondents' creation and their evaluation, the different choice model specifications, the best designs and, stated survey experience from our respondents.

We found that respondents succeeded in covering a large subset of the design space. Namely, we found that they prompted for a diverse set of attributes to be represented on the can. While many gravitated towards more common themes, such as Santa and Christmas trees, many also submitted more individualized and person-specific preferences. Table 2 presents a selection of attributes from each of the three approaches. It emphasizes the different codes as determined by the different methods. Additionally, Figure 10 illustrates six samples that include the attributes from Table 2 in their design. Moreover, it further highlights the diversity, which we already observed in our prompts, but now through the visual medium.

Table 2: A collection of selected attributes from the three different approaches towards coding.

Text Hand Coded	Text AI Coded	Image AI Coded
Stars	Woodland Creatures	Snowflakes
Santa	Wintry Weather	Gifts
Snowman	Red	Moon
Children	Non-traditional color	Reindeers
Animals	Religious theme	Fireplace
Christmas Tree	Elves	Holly berries
Night	Gingerbread house	Mountains
Fireworks	Specific person	Snowy landscape
Manger	Animals besides bears	Text saying Merry Christmas
…	…	…

Figure 10: A selection of images generated by the respondents highlighting the diversity of the produced designs, emphasizing the exploration in our approach.

We additionally found that respondents were generally well-engaged with the exercise. Most adhered to the specified task, submitting valid answers, generating designs that were at least loosely related to Christmas. Only approximately one in five respondent prompts were deemed invalid by our manual evaluation. Only a single respondent was removed altogether from the survey as a result of being flagged by our safety systems.

When coding the images in accordance with our modelling strategy we obtained the results in Table 3.

Table 3: Odds ratios for different codings.

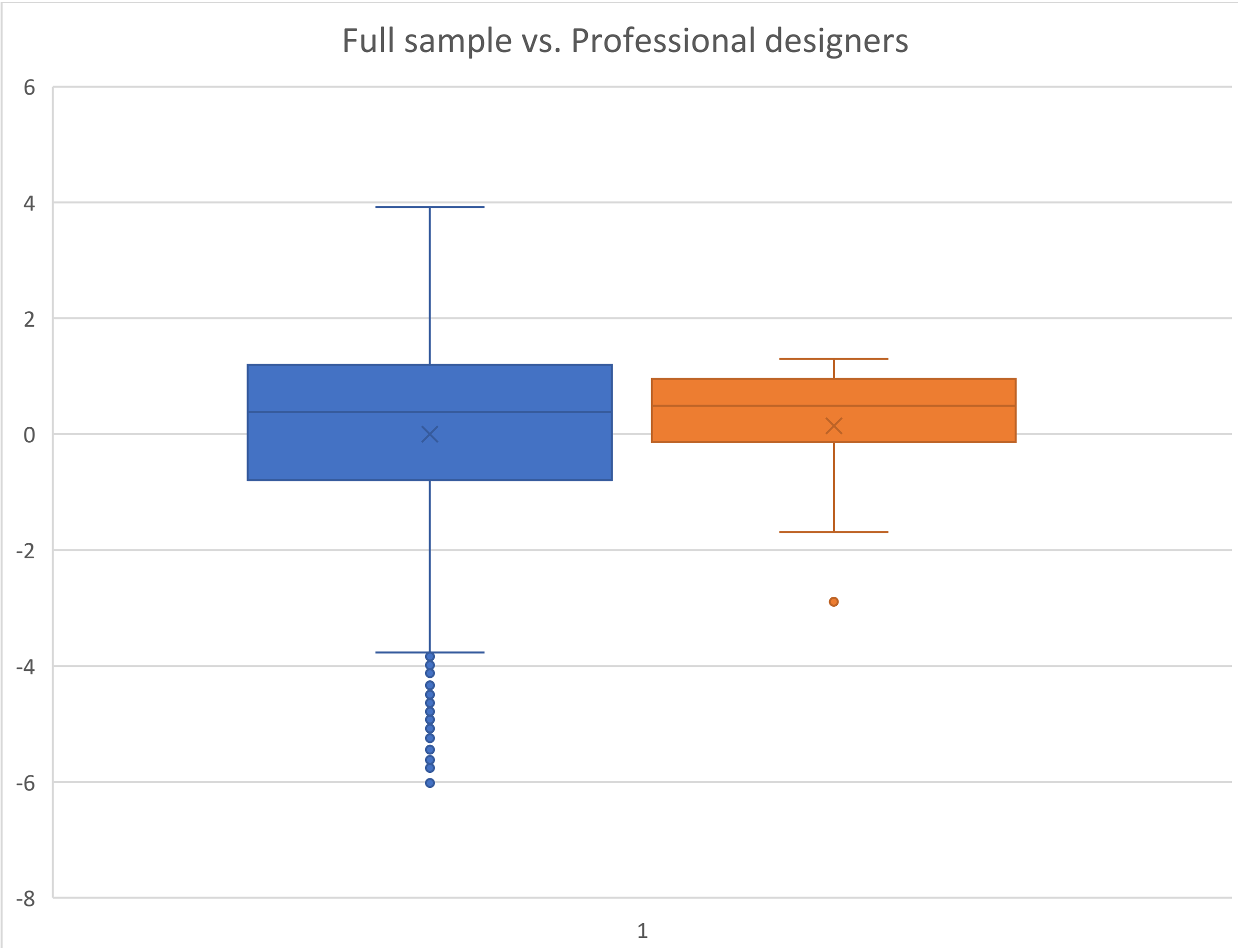
Hand Coded	Odds Ratios	AI coded	Odds Ratios	Image Coded	Odds Ratios
Winter	0.26	Polar bears	0.34	Reindeer	0.37
Animation	0.22	Christmas-themed elements	0.24	Santa Hat	0.27
Animals	0.20	Snowman	0.21	Polar Bear	0.13
Manger	0.20	Santa Suit	0.20	Santa Claus	0.13
Christmas tree	0.18	Holiday Snow Elements	0.19	Snowflakes	0.12
Buildings	0.18	Snow Theme	0.15	Bird	0.11
Bag	0.17	Nativity Scene	0.15	Ornament	0.09
Mr/Mrs Santa	0.17	Winter Wonderland	0.14	Christmas Wreath	0.09

In line with expectations, we achieved the best fit using the Factored Covariance MaxDiff, which allowed us to estimate 1200 concepts and an equivalent number of parameters. This resulted in an RLH of 0.73 and the five cans shown in Figure 11 as the most preferred designs. The top cans displayed marked consistency across the designs, all manifesting attributes from a specific set of attributes. Specifically, all cans featured pine trees and wintery landscapes while additionally exhibiting a similar art style. Nevertheless, each still possessed uniquely distinguishing attributes, for example, the polar bears, wooden cabins, and presents.

Figure 11: The five most preferred designs as determined by our respondent-level MaxDiff.

We included fifteen professional designs in our MaxDiff. Within this survey, they achieved the following rankings: 401, 476, 589, 599, 608, 623, 653, 584, 805, 812, 877, 986, 1070, 1160. We can see this visually in Figure 12, where we see that designers on average performed a little better, but are lacking the well-performing designs that are present in the crowdsourced designs.

Figure 12: Overview of the spacing of utilities of the full sample versus professional designers.



Pertaining to the models we estimated using the hand-coded, AI prompt-coded, and image-coded attributes, we used each task twice. Once positively for best, and then negatively for worst. In addition to estimating the utilities of the attributes, we also estimated a factor as the difference between best and worst, as assuming best is the opposite of worst would be invalid.

Table 4 shows the comparison of the RLHs for the four different models. We can see that the model fit was significantly worse for the attribute-level models. We attribute this performance gap primarily to our modelling choices. Namely, our models first do not consider a difference in generation quality for the same prompt, although we did try to mitigate this issue by generating four images for each prompt. Second, they do not account for the diminishing returns of adding additional features.

Table 4: RLH of the four different models to fit the survey results.

	RLH
MaxDiff*	0.73
Hand Text	0.41
AI Text	0.42
AI Images	0.44

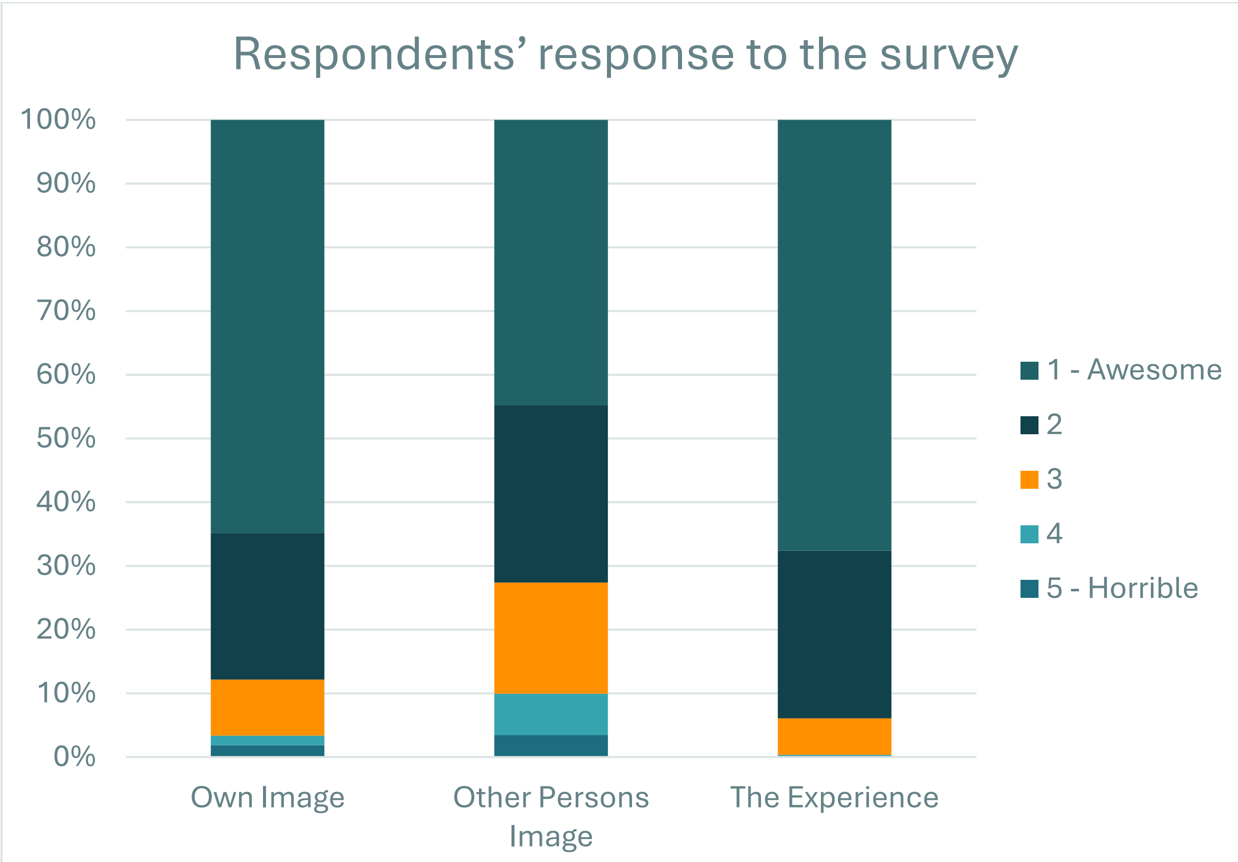
Table 5 presents the correlation of the rankings produced by the four different approaches. When comparing the results we obtain high correlations between the attribute coding approaches, but a lower correlation with the MaxDiff.

Table 5: Correlations between the rankings produced by each of the models.

Correlation	MaxDiff	Hand Text	AI Text	AI Image
MaxDiff	1	0.5	0.5	0.6
Hand Text	0.5	1	0.7	0.6
AI Text	0.5	0.7	1	0.6
AI Images	0.6	0.6	0.6	1

Finally, we asked questions regarding survey enjoyment and design quality. Overall, respondents preferred their own designs over others and rated the survey experience highly when compared to regular surveys. We show the full distribution of ratings in Figure 13.

Figure 13: Respondents' response when asked to rate their own image, other respondents' images, and their overall experience with our survey as compared to regular surveys.



6. Conclusion and Discussion

In this paper, we presented a novel methodology for new product design inspired by crowdsourcing. We conducted a large-scale survey in the US and showed that by using respondents and AI-generated designs, we achieved greater variety, with performance at the upper tail-end of the distribution surpassing professional designs. Additionally, we extracted attributes from our respondent prompts in an automated and scalable fashion using LLMs and used these to construct an interpretable choice model, highlighting the most desirable attributes.

To address present shortcomings, we identify a number of possible avenues for future research. First, from an image generation perspective, we believe improvements with respect to prompt adherence and image fidelity are of paramount importance. A major development that occurred towards the end of our research was the introduction of image editing models, namely GPT-4o from OpenAI (2025) and Flux-Kontext by Black Forest Labs et al. (2025). These models allow us to generate using a base image, ameliorating the need for the brand that we want to generate to be included in the pre-training. These types of models potentially allow us to greatly expand the class of products we can apply our methodology to. For example, we would be able to edit electronic appliances to alter their appearance or add additional features.

Furthermore, from a modeling perspective, we believe that there is still room for improvement with regard to the identification of the features, given the large gap between the respondent- and attribute-level models. We attributed these discrepancies to differences in generation quality, but also the diminishing returns of adding additional features to a single can. We posit that by increasing the complexity of this model, we can achieve better results and more reliable estimates for identifying the winning features.

We would also like to highlight two potential pitfalls. First, at the time of conducting our research, our approach was decidedly novel. Consequently, respondents were unlikely to have participated in something similar before. As a result, the initial high engagement may diminish over time as the approach becomes more prevalent within the market. Nevertheless, we still observe high satisfaction in Figure 13. Finally, pre-prompting in the survey risks restricting the design space to a degree where respondents' creativity is stifled too much, resulting in a loss of design diversity.

Joris van Gool

Peter Li

References

Allen, B., Chandrasekaran, D. & Basuroy, S., 2018. Design crowdsourcing: The impact on new product performance of sourcing design solutions from the "crowd". *Journal of Marketing,* pp. 106–123.

Anon., Radford, Alec and Narasimhan, Karthik and Salimans, Tim and Sutskever, Ilya. *Improving Language Understanding by Generative Pre-Training.* s.l.:s.n.

Black Forest Labs, et al., 2025. *FLUX. 1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space,* s.l.: arXiv preprint arXiv:2506.15742.

Cunningham, C., Deal, K. & Chen, Y., 2010. Adaptive choice-based conjoint analysis: a new patient-centered approach to the assessment of health service preferences. *The Patient: Patient-Centered Outcomes Research,* pp. 257–273.

Gemini, et al., 2023. *Gemini: a family of highly capable multimodal models,* s.l.: arXiv preprint arXiv:2312.11805.

Hartmann, J., Exner, Y. & Domdey, S., 2024. The power of generative marketing: Can generative AI create superhuman visual marketing content?. *International Journal of Research in Marketing,* pp. 13–31.

Ho, J., Jain, A. & Abbeel, P., 2020. *Denoising diffusion probabilistic models.* s.l., s.n., pp. 6840–6851.

Kaplan, J. et al., 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361.*

Lattery, K., Hardt, N. & Huang, H., 2025. *Blending Historical MaxDiff Claim Studies and Using AI to Predict Claim Success.* New Orleans, s.n.

OpenAI, 2025. *Introducing 4o Image Generation,* s.l.: s.n.

Piller, F. T., 2010. Open innovation with customers: crowdsourcing and co-creation at Threadless. *Available at SSRN 1688018.*

Poetz, M. & Schreier, M., 2012. The value of crowdsourcing: can users really compete with professionals in generating new product ideas?. *Journal of product innovation management,* pp. 245–256.

Radford, A., Narasimhan, K., Salimans, T. & Sutskever, T., 2018. Improving Language Understanding by Generative Pre-Training.

Rombach, R. et al., 2022. *High-resolution image synthesis with latent diffusion models.* s.l., s.n., pp. 10684–10695.

Shazeer, N., 2020. *GLU Variants Improve Transformer.* s.l.:s.n.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. & Ganguli, S., 2015. *Deep unsupervised learning using nonequilibrium thermodynamics.* s.l., s.n., pp. 2256–2265.

Zhang, L., Rao, A. & Agrawala, M., 2023. *Adding conditional control to text-to-image diffusion models.* s.l., s.n., pp. 3836–3847.

Better Segmentation Results with Deep Learning: Dimensionality Reduction Using Auto-Encoders

Joseph Retzer
ACT-Solutions

> *"Unsupervised learning is primarily concerned with uncovering latent, non-random structure in data. By uncovering this structure, a deeper understanding of the data, and potentially how it was produced, is possible."*

Philip Waggoner

Abstract

This paper investigates the comparative performance of dimensionality reduction techniques in the context of unsupervised learning, with particular emphasis on cluster analysis. The traditional method of Principal Component Analysis (PCA) (Jolliffe, 2002) is contrasted with deep learning-based Auto-encoders (AEs) (Hinton and Salakhutdinov, 2006), using a dataset of 28 numeric features. Both PCA and AE compress the feature space to seven dimensions before clustering is performed using the Partitioning Around Medoids (PAM) algorithm (Kaufman and Rousseeuw, 1990).

Cluster quality is assessed using the Calin´ski-Harabasz Index (CHI) 1974), Davies-Bouldin Index (DBI) (1979), and Silhouette plots/scores (Rousseeuw, 1987). Results show that auto-encoders substantially outperform both raw data and PCA-transformed data, achieving a silhouette score of 0.32, which exceeds the commonly accepted threshold for interpretability.

In addition to improved clustering results, the paper highlights the interpretability and flexibility of auto-encoder components using correlation heatmaps, dependence plots, and feature permutation importance. While AEs require careful model tuning, their ability to capture nonlinear relationships and preserve more information makes them a superior choice for dimensionality reduction in high-dimensional data. The paper also emphasizes the importance of reproducible feature engineering pipelines and introduces the `tidymodels` ecosystem in `R`, illustrating its use for tasks such as normalization, imputation, and interaction term creation. Overall, the study demonstrates that deep learning methods like auto-encoders can offer meaningful advantages in segmentation and unsupervised modeling tasks.

Preparing Data for Modeling: Feature Engineering

As most researchers are aware, data analysis typically begins with the time-intensive process of transforming raw data into meaningful features that improve model performance. This critical stage, typically the most time-consuming in the modeling workflow, often involves a series of complex and iterative steps.

In the machine learning field, this process is known as "feature engineering." Basic feature engineering encompasses a number of familiar steps outlined below:

- Basic Feature Engineering (data cleaning)
  - Imputing missing data
  - Removing outliers
  - Correcting inconsistencies (spelling errors, duplicate records, formatting inconsistencies, etc.)
  - Converting data types (e.g., dummy/one-hot encoding), etc.

While basic data cleaning is critical for pre-processing data, feature engineering also involves numerous additional transformations, some of which are outlined below:

- Going Beyond Data Cleaning
  - Feature transformation (e.g., normalization, log transformation)
  - Feature creation (creating interaction terms, splines, etc.),—collapsing infrequent levels in high cardinality categorical variables
  - Dimensionality reduction, etc.

Dimensionality reduction, often accomplished through Principal Component Analysis (PCA), is arguably the most critical step when preparing data for cluster analysis. High dimensional data can negatively impact cluster quality and should be dealt with at the outset, before the clustering algorithm is run.

HIGH-DIMENSIONAL DATA: WHAT'S THE PROBLEM?

In order to illustrate the impact of high-dimensional data on cluster analysis results, we may consider the following:

Example: Assume we have 10 variables with an average number of levels equal to 5.

This results in $5^{10} = 9,765,625$ possible combinations.

The feature space in which the data reside is notably large, even for this relatively simple dataset. The resulting data sparsity causes observations to appear nearly equidistant from one another, making it difficult to discern meaningful patterns. While various commonly used clustering algorithms will still produce the requested number of clusters under these conditions, the resulting clusters typically exhibit poor quality due to the lack of distinct groupings.

Moreover, a frequently requested deliverable in applied clustering—namely, a scoring algorithm—is also adversely affected. Specifically, the predictive performance of such algorithms tends to degrade as the distinctiveness of clusters diminishes. This decline in both clustering effectiveness and predictive accuracy, with increasing dimensionality, is a manifestation of the well-known "Curse of Dimensionality." A widely used strategy to mitigate this problem is "Dimensionality Reduction," which may involve reducing the number of variables in a dataset while preserving as much relevant information as possible (e.g., through Principal Component Analysis).

Two approaches to dimensionality reduction, Principal Component Analysis (PCA) and Deep Learning-based Auto-encoders (AE), will be described and illustrated. Cluster analysis results will be compared using data created from each approach along with the raw data.

An overview of the methodology to accomplish this is described below.

- Describe and illustrate PCA and Deep Learning AE's.
- Employ a data set comprised of 28 features, all numeric.
- Use PCA to find optimal number of PC's.
- The same number of Auto-Encoders (AE's) will also be constructed
- Perform cluster analysis using "Partitioning Around Medoids" (PAM) on
    1. Raw Data,
    2. Principal Component Data, and
    3. Auto-Encoder Data using the following metrics:

Resulting partitions will be compared based on:

- The Calin´ski-Harabasz Index (CHI),
- Davies-Bouldin Index (DBI) and,
- Silhouette Value/Plot.

PRINCIPAL COMPONENT ANALYSIS (PCA):

Principal Component's (PC's) are *linear combinations* of the original variables. Weights used for averaging the variables are taken from the eigenvectors of the decomposition of the data covariance matrix. It is straightforward to show that using the eigenvectors as weights results in the maximum variance of the average.

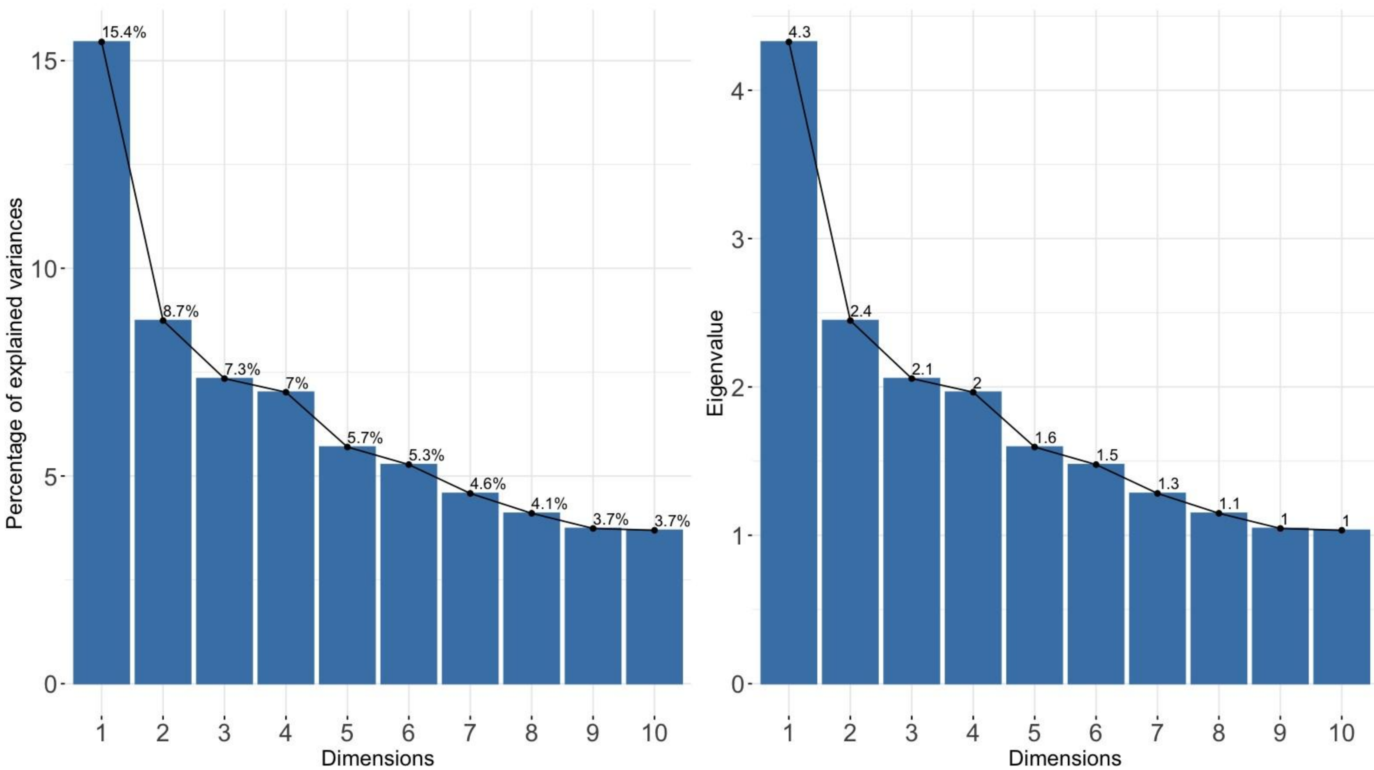
PC's are then ordered by the amount of variance captured from data. Only the first few, most informative PC's are retained. The process therefore reduces the number of dimensions while attempting to retain as much of the data's variability as possible.

Note: Dimensionality reduction using PCA "requires" loss of information since some PC's must be dropped.

PCA Output

A graphical depiction of both the percentage of information and its proxy, eigenvalues associated with each PC, is shown in Figure 1 below.

Figure 1: PCA Output

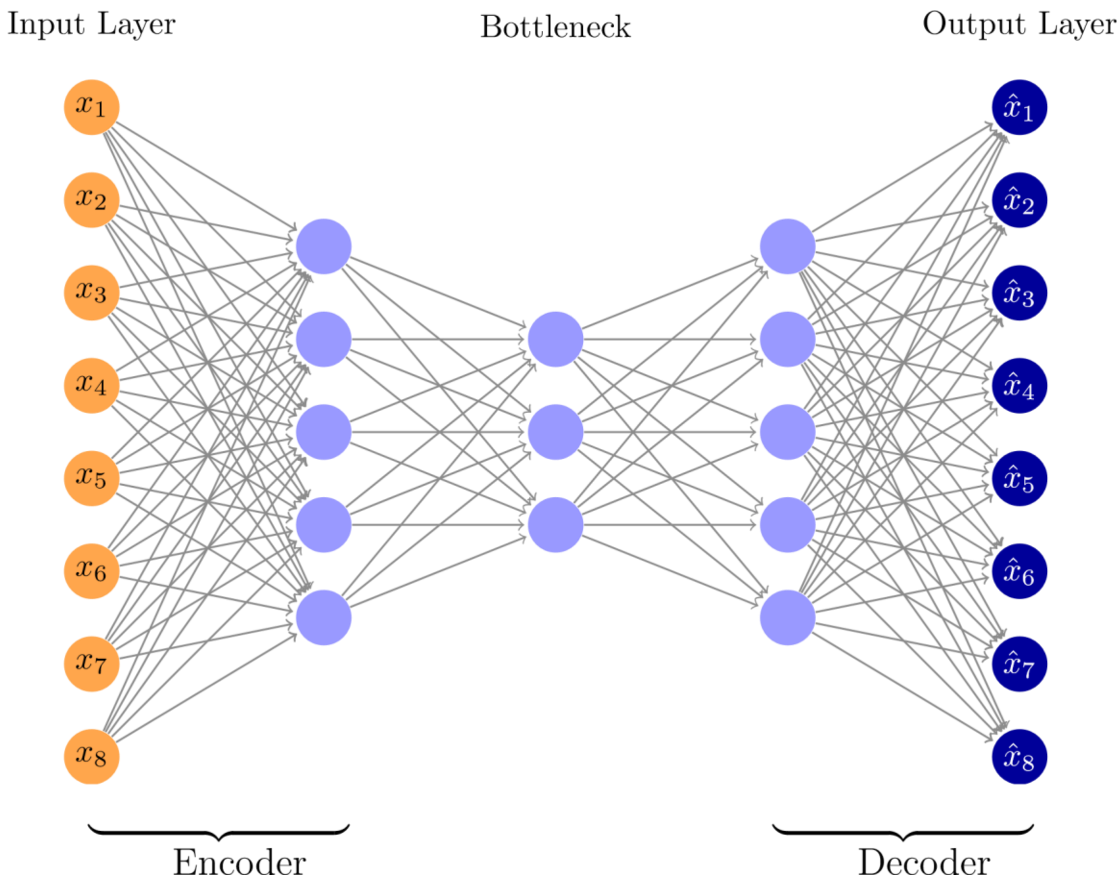


Based on the PCA output, the first 7 PC's were chosen to replace the data used to construct them. The same number of Auto-Encoders will also be selected.

Neural Network Based Auto-Encoders

A graphical depiction of a simple auto-encoder is shown in Figure 2 below.

Figure 2: Simple Auto-Encoder



Note that each column of circles is referred to as a "layer" and each circle in a column is referred to as a "node."

The auto-encoder depicted in Figure 2 illustrates a **fully connected, symmetric auto-encoder** with the following structure:

- **Input Layer:** Consists of 8 input features (nodes), denoted as $x_1,x_2,\ldots,x_8$. This is the raw data.
- **Bottleneck Layer:** A central latent space where the input data is compressed into a lower-dimensional representation (3 auto-encoders in this example) known as the bottleneck layer. This layer captures the most salient features of the input.
- **Output Layer:** Comprises 8 output nodes, labeled $\hat{x}_1,\hat{x}_2,\ldots,\hat{x}_8$, which aim to reconstruct the original data.

This auto-encoder is designed to learn a compact encoding of 8-dimensional input data via the bottleneck layer. The encoder compresses the input into a lower-dimensional latent representation, while the decoder reconstructs the input from this compressed form.

The network is trained to minimize reconstruction error—in this case it would be the mean squared error (MSE) loss function involving the difference between the actual and reconstructed features—thereby enabling the discovery of efficient, information-preserving representations of the original data.

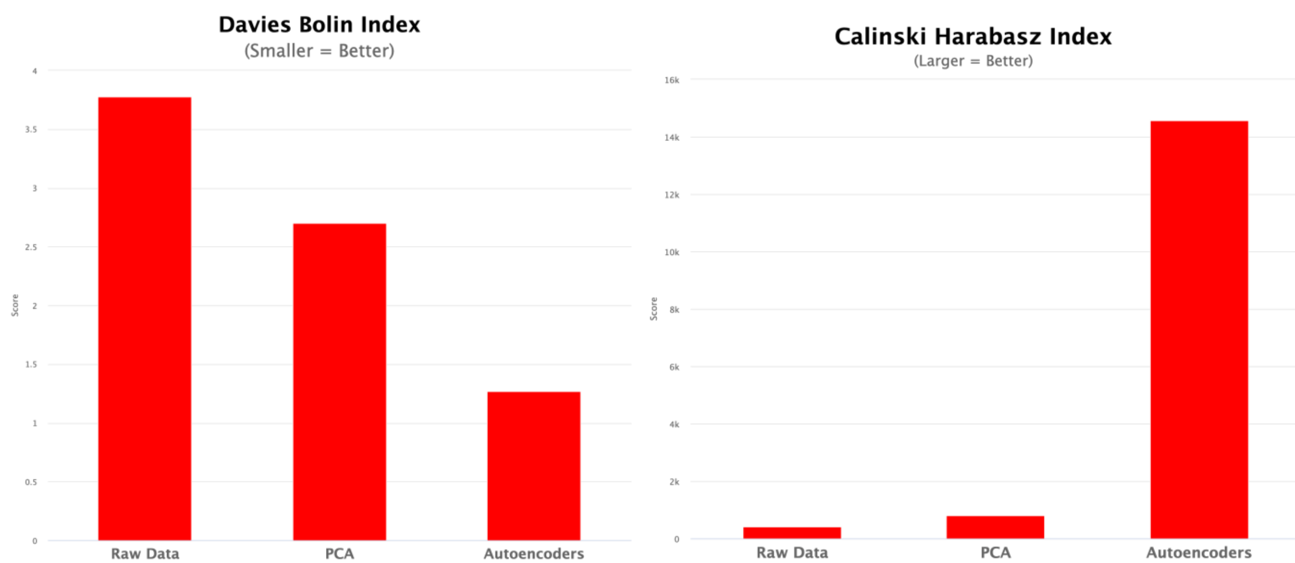
Comparing Raw Data, PC’s and AE using Cluster Analysis

As noted earlier, partition quality metrics **Calin´ski-Harabasz Index (CHI), Davies-Bouldin index (DBI)** and the **Silhouette Score/Plot** will be used to evaluate cluster solutions resulting from “Partitioning Around Medoids” (PAM) cluster analysis. The CHI and DBI are both internal cluster validity indices used to evaluate clustering quality, however CHI is often preferred for evaluating well-separated and compact clusters, particularly in high-dimensional settings.

The **Silhouette Score** measures how well each data point fits within its assigned cluster. It also provides a graphical depiction of quality by cluster.

As evidenced in the graphics shown in Figure 3, both DBI and CHI strongly indicate a higher quality solution provided using auto-encoders.

Figure 3: DBI and CHI Metrics



Next, the silhouette plots (and average silhouette scores) will be compared again using partitions derived from (1) raw data, (2) PC’s and (3) auto-encoders.

As shown in the silhouette plots in Figure 4, both the clustering partitions based on the raw data and those derived from the principal components exhibit unacceptably low average silhouette values. Neither partition achieves the commonly accepted threshold of 0.2, which is typically considered the minimum for an interpretable/reliable solution.

Figure 4: Silhouette Plots: Raw Data and PC's

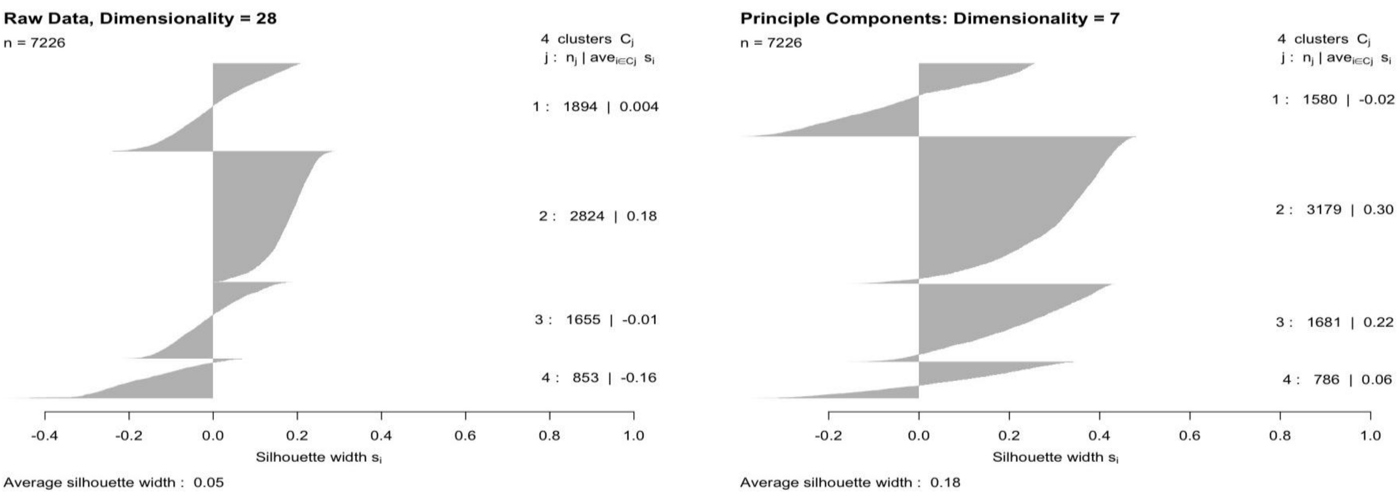
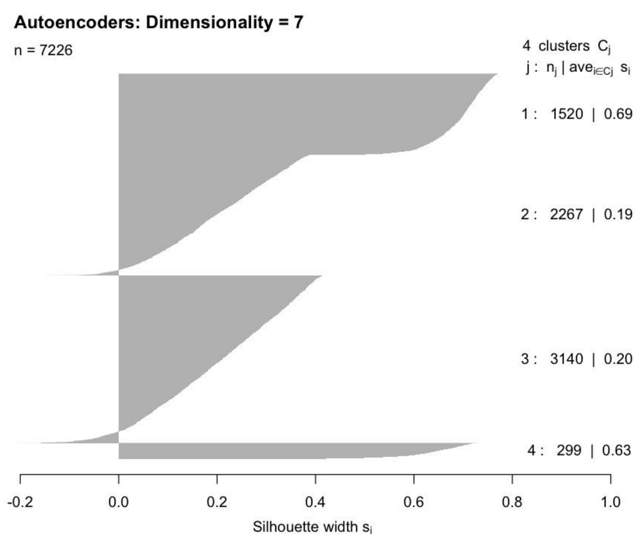


Figure 5 reflects the silhouette plot when using auto-encoders as data. This plot clearly indicates a greatly improved partition with average silhouette calculated as .32, well above the minimal threshold of .2.

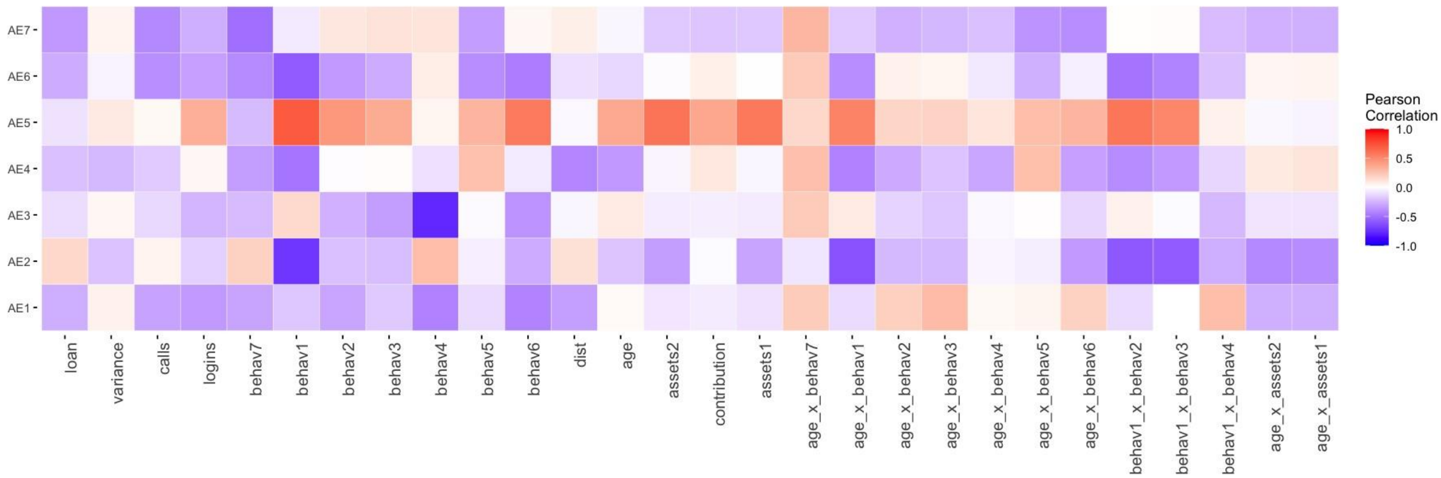
Figure 5: Auto-Encoder Silhouette: Average Silhouette = .32



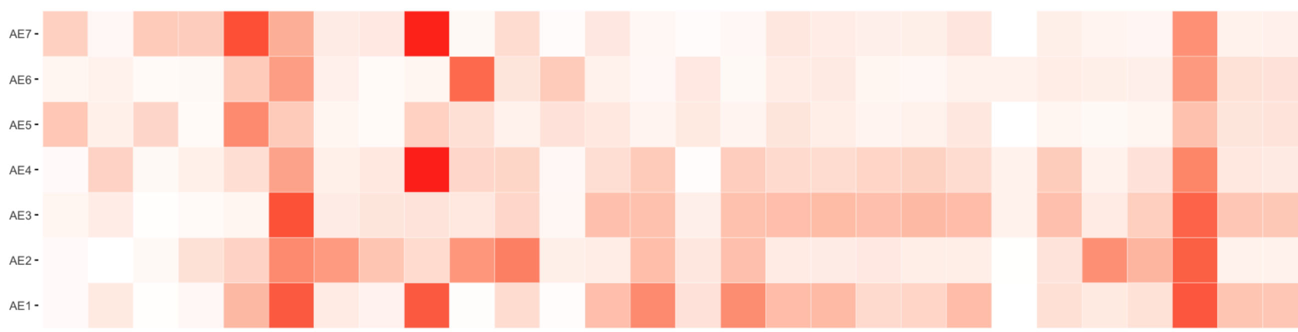
Given that the auto-encoders yield higher-quality partitions, it is important to interpret the structure and meaning of each resulting component. To facilitate this understanding, a range of visualization techniques may be employed. Among them, a correlation heatmap and a more general dependence measure heatmap—such as the one proposed by Chatterjee (2021)—are particularly useful. These visualizations are presented in Figures 6 and 7. In these plots, the vertical axis corresponds to the auto-encoder components, while the horizontal axis represents the original input features.

Figure 6: Correlation and Dependence Heat Maps

Correlation Heatmap



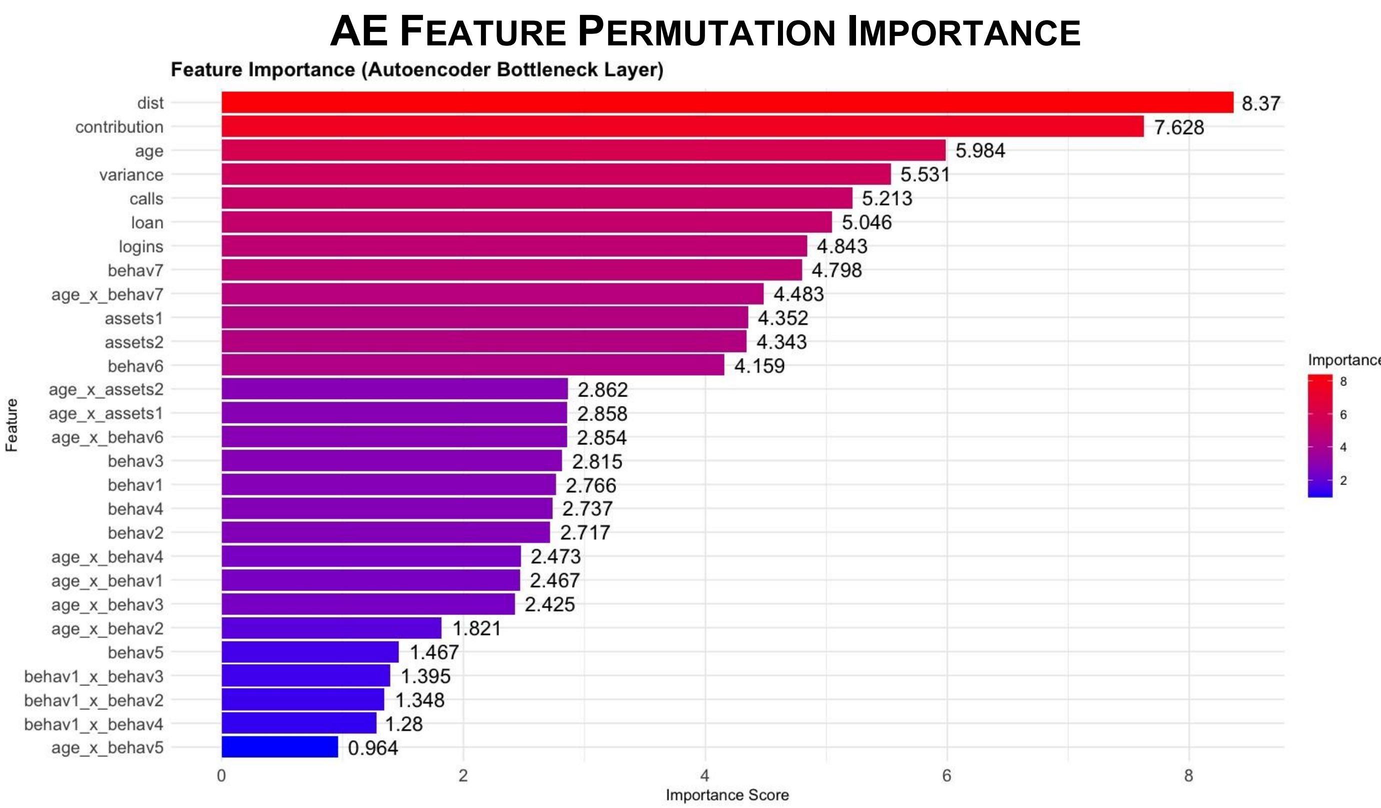
Dependence Heatmap



See S. Chatterjee, A New Coefficient of Correlation (2020), JASA.

Another potentially useful visualization pertaining to auto-encoders is "Feature Permutation Importance Bar Chart." This involves making comparisons between approximated feature values and the original data before and after permuting a specific feature of interest. The greater the difference in actual vs. estimated features due to the permutation, the more important is the feature in the construction of the auto-encoders. A feature permutation bar chart is illustrated in Figure 7.

Figure 7: Feature Permutation Importance



Based on this example, several conclusions can be drawn regarding the comparative effectiveness of auto-encoders and principal component analysis (PCA) as techniques for dimensionality reduction:

- **Dimensionality Reduction**
  - AE's learn non-linear relationships and preserve information. This may result in higher quality cluster solutions.
  - It is also noteworthy that, due to their superior ability to retain information, auto-encoders may offer more effective two- or three-dimensional graphical representations of the data compared to principal components.

- AE's may also be used to detect data anomalies, as well as de-noising data for subsequent analysis.

- **Drawback of AE** Unlike PCA, auto-encoder construction may involve significant model tuning. Numerous measures may be tuned including, but not limited to: gradient descent learning rate, optimal number of nodes, layers, epochs, batch size (number of observations used in a single forward/backward propagation. Default is usually 32), activation function(s) choice, etc.

Feature Engineering with tidymodels

While dimensionality reduction is often the most critical feature engineering step in cluster analysis, it is typically accompanied by a range of additional data cleaning and transformation tasks. An effective suite of tools for carrying out these operations is available within the tidymodels ecosystem in R. This collection of packages not only facilitates the easy implementation of various feature engineering techniques but also enables the construction of pipe-able/reproducible workflows that can be applied consistently across multiple datasets.

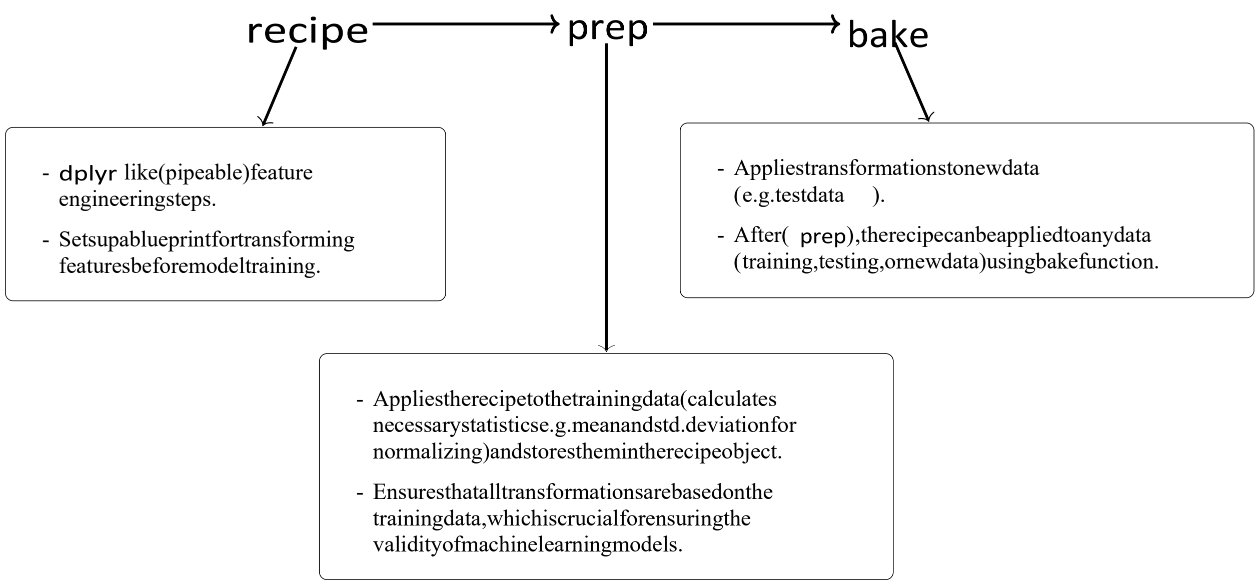
More specifically, tidymodels is a unified framework in R that provides a coherent, modular, and reproducible approach to data preprocessing, model development, evaluation, and tuning. Its consistent syntax and design philosophy promotes transparency and scalability in the modeling process.

Three functions for data pre-processing are particularly useful for this purpose. They are:

- recipe
- prep
- bake

These functions, along with descriptions, are illustrated in the diagram in Figure 8 below:

Figure 8: tidymodels: Recipe, Prep, Bake



An example of a recipe is shown below. This recipe takes a dataset called myDat and performs the transformations noted in each step:

recipe **Example** myData recipe <- recipe(id ~ .,

```
data=myDat) %>%

# Impute missing values with KNN
   step_impute_knn(all_predictors(),neighbors=5) %>%
```

```
# Create interaction terms step_interact(terms =~
   age:starts_with("behav")) %>%

# Group infrequent levels: "other" step_other(myCatVar,
   threshold = 0.001) %>%

# Approximates normalizing transformation
   step_orderNorm(all_numeric_predictors()) %>%

# PCA step_pca(all_numeric_predictors(), num_comp = 7)
```

Numerous additional steps are available including

- step downsample: balance a model response variable for predictive modeling,
- step spline natural: create splines for non-linear feature modeling,
- step zv: identify and remove features with zero variance, etc., as just a few additional examples.

CONCLUSIONS

For the data examined, this work demonstrated how auto-encoders outperformed both PCA and raw data in terms of producing high quality clusters evidenced by all three metrics used: DBI, CHI and Silhouette Scores. Specifically, AE-based data produced a significantly higher average silhouette score (0.32), exceeding the minimum acceptable threshold (0.2), unlike PCA and raw data partitions.

Visualization tools—including correlation and dependence heatmaps and feature permutation importance plots—provided interpretability of the AE components.

Several clear advantages of auto-encoders were identified/noted:

- AE's can learn non-linear relationships and preserve more information than PCA.
- They enable more informative low-dimensional visualizations.
- AE's can also be used for anomaly detection and de-noising.

One caution regarding AE implementation involves model tuning (e.g., learning rate, architecture, activation functions), making it more complex than PCA. However, it seems the advantages far outweigh the drawbacks of using auto-encoders for dimensionality reduction.

In addition to using auto-encoders for dimensionality reduction, tools for implementing addition feature engineering steps were also reviewed. Specifically, the tidymodels framework in R is highlighted as a powerful and reproducible approach to feature engineering and model development. Key components of this ecosystem include functions: recipe(), prep(), and bake() which streamline transformation pipelines, supporting tasks like imputation, interaction term creation, normalization, and PCA to name but a few.

This work demonstrates that deep learning methods, specifically auto-encoders, offer a powerful enhancement to traditional dimensionality reduction techniques and can lead to significantly better segmentation and modeling outcomes in unsupervised learning applications.

Joseph Retzer

References

Calin´ski, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27.

Chatterjee, S. (2021). A new coefficient of correlation. *Journal of the American Statistical Association*, 116(536):2009–2022.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. In *Science*, volume 313, pages 504–507. American Association for the Advancement of Science.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer Series in Statistics. Springer, 2nd edition.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Statistics. John Wiley and Sons.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

ENHANCING CLUSTER ENSEMBLES WITH LATENT CLASS CLUSTERING

KEITH CHRZAN
SAWTOOTH
JOSEPH WHITE
KYNETEC

Strehl and Ghosh (2002) introduced cluster ensembles—the idea of aggregating multiple clustering solutions to arrive at a consensus solution. Retzer and Shan (2007) brought these ideas to the Sawtooth community, prompting Orme and Johnson (2008) to incorporate ensemble analysis into Sawtooth's CCEA module. CCEA's default ensemble consists of 70 clustering solutions across 14 predefined segment sizes ranging from 2 to 30. Each segment size includes five different algorithms or initializations: k-means with distance-based seeds, density-based seeds, seeds derived from hierarchical clustering, and hierarchical clustering using both complete and average linkage. These solutions are then recoded into binary membership indicators and subjected to convergent k-means meta-clustering to produce a final consensus segmentation.

CCEA's default of building its ensemble from different numbers of segments and different clustering algorithms could be expanded in various ways, like looking at subsets of variables. Below we explore whether augmenting the CCEA ensemble with latent class clustering solutions can improve segmentation outcomes. We can use model-based approaches such as R's mclust package or commercial software like Latent Gold to run our latent class clustering (AKA model-based clustering or finite mixture modeling). By assuming that data arise from a mixture of underlying probability distributions, these models provide a statistically principled way of uncovering latent group structure.

In previous research (Chrzan and White, 2021) we found that model-based clustering methods like mclust perform comparably to CCEA in terms of identifying the correct number of clusters and achieving accurate respondent classification. Given this, it is reasonable to investigate whether incorporating model-based latent class clustering solutions into the ensemble can further improve performance. In this paper we use 47 artificial data sets varying different segmentation-relevant characteristics. Having imposed the segment structure in each of these data sets, we seek to learn whether integrating model-based methods into the CCEA ensemble improves our ability to identify the correct number of segments and our ability to assign the right respondents to the right segments.

We plan to compare CCEA with default setting, model-based clustering with default settings, and CCEA that augments the default ensemble with model-based clustering solutions. We evaluate the performance of these three methods using two primary metrics. First, we examine how accurately each method recovered the correct number of segments. For model-based methods, we used the solution with the lowest Bayesian Information Criterion (BIC), while for ensemble methods, we used the solution associated with the largest uptick in the CCEA reproducibility statistic. Second, we assess classification accuracy using the Adjusted Rand Index (ARI), which compares estimated segment assignments with known true memberships. We calculate the ARI based on a method's solution for the correct number of segments, regardless of whether a method identified that number correctly, because we wanted to separate the success of the methods in terms of identifying the correct number of segments and putting the right respondents in the right segments. The ARI usually ranges from 0 (random agreement) to 1 (perfect agreement), correcting for chance matches based on pairwise comparisons (Morey and Agresti, 1984). For example, a perfect ARI of 1.00 might look like this:

	1	**2**	**3**	**4**
1	89	0	0	0
2	0	33	0	0
3	0	0	75	0
4	0	0	0	112

With one segmentation's assignments in the banner and another's in the stub, a perfect match would have all respondents on the diagonal and all off-diagonal cells empty (if, as we have here, rows and columns are sorted to be in the same order).

With an ARI of 0.71, things get a little messier, with more respondents falling into the off-diagonal cells. For example:

	1	**2**	**3**	**4**
1	140	8	0	2
2	1	148	35	0
3	9	5	176	0
4	6	0	0	54

Finally, an ARI of 0.33 is messier still, as in:

	1	2	3	4
1	120	50	10	20
2	40	110	30	20
3	10	30	130	30
4	30	10	40	140

Analysis and Results

Baseline and Treatments

Five clustering strategies are initially considered in evaluating the potential benefit of enhancing the default CCEA ensemble with latent class solutions. These are comprised of three base cases and two treatment cells. Base cases include CCEA, Latent Gold, and mclust, each using its own default settings. These base cases include ensembles of size 70, 1, and 1, respectively. The treatment cells expand the default CCEA ensemble with an additional set of 14 solutions consistent with the number of segments in the default ensemble. The treatment cells then have 84 cluster solutions each.

Study 1 Data Sets

The initial data sets analyzed are those from Sawtooth's technical paper (Orme and Johnson, 2008) evaluating the potential of CCEA over Convergent Cluster Analysis (CCA). That paper considers eleven synthetic data sets differing by number of segments, number of basis variables, and segment size as specified in the following table.

Data Set	Segments	Basis Variables	Segment Size Ratio
1	3	10	1:3:6
2	3	10	2:3:5
3	3	10	1:1:1
4	3	10	1:3:6
5	3	10	1:3:6
6	4	25	1:2:3:4
7	6	10	1:2:3:4:4:6
8	6	10	1:2:3:4:4:6
9	6	10	1:2:3:4:4:6
10	6	10	1:2:3:4:4:6
11	6	10	1:2:3:4:4:6

Data set 6 consists of 500 records compared to 1,000 for the other 10 data sets. Data sets 7 through 11 have increasing within-segment standard deviations (11 the largest). See Orme and Johnson, 2008, for further details on data set construction.

Study 1 Results

The five methods (baselines and treatments) are evaluated based on their ability to identify the correct number of segments and assignment of records. The number of clusters identified by method for each data set is reported in the following table. The mean absolute deviation (MAD) summarizes the performance across data sets for each clustering strategy, with lower values being better.

	Number of Segments					
Data set	**TRUE**	**mclust**	**Latent Gold**	**CCEA**	**CCEA+ mclust**	**CCEA+LG**
1	**3**	3	2	2	2	4
2	**3**	3	3	2	2	2
3	**3**	3	3	3	3	3
4	**3**	3	2	3	3	3
5	**3**	3	2	2	2	2
6	**4**	4	5	3	4	5
7	**6**	6	6	6	6	6
8	**6**	6	6	6	6	5
9	**6**	5	5	5	5	5
10	**6**	5	5	5	5	5
11	**6**	5	2	5	5	5
	MAD Sets 1-11	***0.27***	***0.91***	***0.64***	***0.55***	***0.73***
	MAD Sets 1-10	***0.20***	***0.60***	***0.60***	***0.50***	***0.70***

The green shaded cells indicate the correct number of clusters is identified, the peach cells where the identified number of clusters is 1 off the true number, and the red cell off by 4. The mclust enhanced ensemble outperforms default CCEA but neither performs as well as the mclust base case. Excluding data set 11, Latent Gold (LG) and CCEA base cases are at parity while the LG enhanced ensemble performs worse than either.

Turning to the correct assignment of records to their true segment, we consider the ARI as described earlier. As a reminder, a higher ARI reflects better classification. ARI detail results are presented below.

Data set	ARI Between Solution and True Membership				
	mclust	Latent Gold	CCEA	CCEA+ mclust	CCEA+LG
1	0.71	0.71	0.64	0.68	0.68
2	0.51	0.49	0.44	0.46	0.49
3	0.43	0.43	0.44	0.44	0.44
4	0.50	0.44	0.34	0.37	0.37
5	0.81	0.79	0.65	0.68	0.68
6	0.87	0.77	0.89	0.89	0.89
7	1.00	1.00	1.00	1.00	1.00
8	0.99	0.99	0.97	0.99	0.98
9	0.83	0.83	0.66	0.68	0.68
10	0.50	0.56	0.55	0.58	0.56
11	0.42	0.33	0.35	0.37	0.35
Mean Sets 1-11	***0.69***	***0.67***	***0.63***	***0.65***	***0.65***
Mean Sets 1-10	***0.72***	***0.70***	***0.66***	***0.68***	***0.68***

Recall that data sets 7 through 11 differed by within segment standard deviation, with 7 having the lowest and 11 the highest. The increasing within-segment variation presents a challenge to all methods as evidenced by the declining ARI over those data sets. Again, the mclust base case outperforms all other methods, and both mclust and Latent Gold as a standalone solution do a better job at classifying records than CCEA or either of the enhanced ensemble strategies.

At this point, we also calculated the ARI between mclust and Latent Gold for each of the 11 data sets. The average ARI for this analysis was 0.855, indicating the two model-based clustering approaches produce similar results. Due to this similarity and the relative ease of automation in R, we focus only on mclust for the latent class solutions in the remainder of the paper.

Studies 2–4 Experimental Design

Results from Study 1 suggest that model-based clustering techniques can help improve CCEA through an enhanced ensemble but also that mclust as a standalone solution may be preferred. We expand on the analysis with three rounds of new synthetic data sets that we refer to as Studies 2, 3, and 4. Each study included 12 separate data sets, structured according to an experimental design varying the number of segments, segment shape, and relative segment sizes, with details provided in the following table.

Data Set	Segments	Shape	Segment Sizes
1	3	Spherical	350:350:350
2	3	Ellipsoidal	350:350:350
3	3	Spherical	600:300:150
4	3	Ellipsoidal	600:300:150
5	5	Spherical	227:227:227:227:227
6	5	Ellipsoidal	227:227:227:227:227
7	5	Spherical	450:290:190:125:80
8	5	Ellipsoidal	450:290:190:125:80
9	7	Spherical	169:169:169:169:168:168:168
10	7	Ellipsoidal	169:169:169:169:168:168:168
11	7	Spherical	300:240:190:150:120:100:80
12	7	Ellipsoidal	300:240:190:150:120:100:80

Previous research (Chrzan and White 2022) showed the importance of preliminary variable selection to remove redundant and masking variables, so we do not include those as part of the experimental design. All data sets have 8 basis variables.

Data Generation Process

The data generation process starts with the construction of a set of segment centroids resulting from random draws from a source distribution. Depending on the study, the source distribution is either multivariate normal with mean 0, variance 2, and covariance 0 between dimensions or uniform with a min of 1 and max of 10. Taking random draws means there are no proximal restrictions on centroids in the data generation process, i.e., cluster centroids can be very far or very near one another for any given dimension.

Random perturbations are then added to the centroid for the cluster associated with each record. Depending on the study, these perturbations are taken from multivariate normal or normal distributions with mean 0 and shape appropriate scale parameters. Hyperspherical clusters take perturbations from a distribution with constant variance, whereas ellipsoidal clusters are generated using distributions with dimension varying scale parameters.

Depending on the study, we either allow or disallow segments to overlap from a distance perspective. When overlap is disallowed, we require that all members be closer to their own shifted centroid than any other. An iterative process replaces overlapping records with new candidates until there is no overlap. When overlap is allowed, the process ends after adding the perturbations regardless of centroid proximity.

The following table presents the study specific data set parameters.

<table>
<tr><th></th><th>Study 2</th><th>Study 3</th><th>Study 4</th></tr>
<tr><td>Centroids</td><td colspan="2">Multivariate Normal
Mean = 0, VAR = 2, COV = 0</td><td>Uniform
Min = 1, Max = 10</td></tr>
<tr><td>Overlap</td><td>No</td><td>Yes</td><td>Yes</td></tr>
<tr><td>Hypersphere</td><td colspan="2">Multivariate Normal
Mean = 0, VAR = 2, COV = 0.2</td><td>Normal
Mean = 0, VAR = 3</td></tr>
<tr><td>Ellipsoidal</td><td>Multivariate Normal
Mean = 0, COV = 0.2
VAR = 1 to 6.25 by 0.75</td><td>Multivariate Normal
Mean = 0, COV = 0.2
VAR
<table>
<tr><th></th><th>X1</th><th>X2</th><th>X3</th><th>X4</th><th>X5</th><th>X5</th><th>X7</th><th>X8</th></tr>
<tr><td>seg 1</td><td>1</td><td>2</td><td>3</td><td>2</td><td>1</td><td>2</td><td>1</td><td>3</td></tr>
<tr><td>seg 2</td><td>3</td><td>1</td><td>2</td><td>3</td><td>2</td><td>1</td><td>2</td><td>1</td></tr>
<tr><td>seg 3</td><td>1</td><td>3</td><td>1</td><td>2</td><td>3</td><td>2</td><td>1</td><td>2</td></tr>
<tr><td>seg 4</td><td>2</td><td>1</td><td>3</td><td>1</td><td>2</td><td>3</td><td>2</td><td>1</td></tr>
<tr><td>seg 5</td><td>1</td><td>2</td><td>1</td><td>3</td><td>1</td><td>2</td><td>3</td><td>2</td></tr>
<tr><td>seg 6</td><td>2</td><td>1</td><td>2</td><td>1</td><td>3</td><td>1</td><td>2</td><td>3</td></tr>
<tr><td>seg 7</td><td>3</td><td>2</td><td>1</td><td>2</td><td>1</td><td>2</td><td>1</td><td>2</td></tr>
</table></td><td>Normal
Mean = 0
VAR
<table>
<tr><th></th><th>X1</th><th>X2</th><th>X3</th><th>X4</th><th>X5</th><th>X5</th><th>X7</th><th>X8</th></tr>
<tr><td>seg 1</td><td>1</td><td>4</td><td>9</td><td>4</td><td>1</td><td>4</td><td>1</td><td>9</td></tr>
<tr><td>seg 2</td><td>9</td><td>1</td><td>4</td><td>9</td><td>4</td><td>1</td><td>4</td><td>1</td></tr>
<tr><td>seg 3</td><td>1</td><td>9</td><td>1</td><td>4</td><td>9</td><td>4</td><td>1</td><td>4</td></tr>
<tr><td>seg 4</td><td>4</td><td>1</td><td>9</td><td>1</td><td>4</td><td>9</td><td>4</td><td>1</td></tr>
<tr><td>seg 5</td><td>1</td><td>4</td><td>1</td><td>9</td><td>1</td><td>4</td><td>9</td><td>4</td></tr>
<tr><td>seg 6</td><td>4</td><td>1</td><td>4</td><td>1</td><td>9</td><td>1</td><td>4</td><td>9</td></tr>
<tr><td>seg 7</td><td>9</td><td>4</td><td>1</td><td>4</td><td>1</td><td>4</td><td>1</td><td>4</td></tr>
</table></td></tr>
</table>

Ellipsoidal segments in Study 2 are generated by applying dimension level variances as opposed to Studies 3 and 4 that implement segment-dimension level variances as specified in the sub-tables. Study 3 differs from Study 2 further by the allowance of overlap. The main difference between Study 3 and Study 4, other than the initial centroid source distribution, is that Study 4 data is constructed using a series of independent draws from a univariate normal distribution rather than single draws from a multivariate normal. Thus, any correlation between dimensions is due to chance rather than construction as with Studies 2 and 3.

RESULTS

The addition of these 36 data sets gives us a total of 47 for our analyses. This section presents the results across all data sets from each of the studies. However, as noted earlier we exclude the Latent Gold solutions from the analyses.

We first look at success in identifying the correct number of segments for the base cases of default CCEA and mclust. The following contingency table presents head-to-head results.

<table>
<tr><th></th><th></th><th colspan="5">CCEA Size of Miss</th><th></th></tr>
<tr><th rowspan="6">mclust Size of Miss</th><th></th><th>0</th><th>1</th><th>2</th><th>3</th><th>4</th><th rowspan="6">mclust Better</th></tr>
<tr><th>0</th><td>13</td><td>6</td><td>1</td><td>2</td><td>0</td></tr>
<tr><th>1</th><td>2</td><td>10</td><td>3</td><td>1</td><td>1</td></tr>
<tr><th>2</th><td>2</td><td>1</td><td>3</td><td>1</td><td>0</td></tr>
<tr><th>3</th><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><th>4</th><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
</table>

CCEA Better

Row entries show the size of miss for CCEA conditional on the size of miss for mclust, and vice versa for columns. Table entries are data set counts. The diagonal shows the number of times CCEA and mclust performed equally well or equally poorly for a given data set. For example, for 13 data sets, both CCEA and mclust identified the correct number of clusters. mclust identified the correct number of clusters in 22 (13+6+1+2+0) data sets. Entries above the diagonal represent data sets where mclust outperformed CCEA, and those below where CCEA outperformed mclust. In 15 of the 47 cases, mclust outperformed CCEA, whereas the converse only occurred 6 times.

The research question however, is whether adding mclust solutions to the default ensemble of CCEA improves clustering results. The next set of tables compares performance of the enhanced ensemble to the default for both CCEA and mclust.

CCEA Size of Miss	CCEA + mclust Size of Miss: 0	1	2	3	4
0	18	0	0	0	0
1	3	12	2	0	0
2	2	0	5	0	0
3	2	0	0	2	0
4	0	0	0	0	1

mclust Size of Miss	CCEA + mclust Size of Miss: 0	1	2	3	4
0	17	3	1	1	0
1	3	9	3	1	1
2	4	0	3	0	0
3	1	0	0	0	0
4	0	0	0	0	0

Adding mclust solutions to the ensemble does improve performance for CCEA, doing better than the default in 7 data sets compared to 2. It is noteworthy that in each of the 7 data sets where the enhanced ensemble outperforms the default, the enhanced ensemble identifies the correct number of segments. This indicates that the addition of latent class solutions improves CCEA. However, we also see that mclust alone outperforms the enhanced ensemble in 10 cases compared to 8, suggesting it may be preferred to generating the enhanced ensemble.

The MAD across data sets better presents the story. Adding mclust provides a nice improvement to CCEA but is still not as good as mclust alone.

	Mclust	CCEA	CCEA+mclust
Mean Absolute Deviation	0.72	1.00	0.77

Finally, we consider the ARI and how well our clustering strategies correctly classify records. Results are reported in the table below.

	Adjusted Rand Index Summary		
	mclust	**CCEA**	**CCEA+mclust**
Mean	0.671	0.654	0.687
Wins/Ties (# max)	27	8	15
> mclust	-	18	20
> CCEA	28	-	32
> CCEA+mclust	25	14	-

Overall, based solely on the average ARI across data sets, the enhanced ensemble outperforms both mclust and CCEA, although only marginally besting mclust. Given Study 1's findings and the previous discussion on identifying the correct number of clusters, this is not exactly what we expect to see. If mclust enhanced solutions benefit CCEA but outperform the enhanced ensemble as a standalone strategy, we would expect to see the CCEA+mclust ARI fall between that of mclust and CCEA. This leads us to a deeper dive into respondent classification.

In light of this apparent disconnect, we tally comparative performance at the data set level. The "Wins/Ties (# max)" row in the table above shows for how many data sets each strategy resulted in the best record classification, which includes ties. The story that emerges here is more consistent with what we see in the MAD numbers. Although mclust is middle of the road based on an overall average, it dominates in terms of consistently having the best ARI among the three. In 27 of the 47 data sets, mclust alone results in the best fit, whereas the same can be said for CCEA in only 8 data sets and CCEA+mclust in 15. That the enhanced ensemble improves upon CCEA but falls short of mclust is better aligned with what we would have expected.

The bottom three rows in the table report the pairwise comparison results. Row entries indicate the number of data sets for which the column strategy bests the row strategy. For example, CCEA outperforms mclust on ARI 18 times and CCEA+mclust bests mclust alone 20 times. Because the entries in the last 3 rows are strict pairwise comparisons, they can only be less than the number of wins/ties by the number of ties. In particular, mclust is tied with CCEA in one data set and CCEA+mclust in 2 data sets, all of which have an ARI of 1. A larger discrepancy between the pairwise and three-way (wins/ties row) comparisons means a greater tendency to fall in the middle. We see that mclust tends to be either the worst or the best, and CCEA and CCEA+mclust tend to be more in the middle. This explains why we see the relative ranking at the mean ARI level and leaves our findings intact.

Conclusion

We've seen that augmenting CCEA's default ensemble with model-based solutions improves its performance. But model-based clustering alone usually outperforms CCEA with a default ensemble or with a model-based augment to the default ensemble. This suggests that latent class clustering—whether through open-source packages like mclust or commercial software such as Latent Gold—should be considered a primary segmentation approach in applied segmentation research.

Analysts using CCEA should consider incorporating latent class solutions into their ensembles when feasible, particularly if they suspect their data exhibit overlapping clusters or heteroskedasticity. However, in many cases, a direct application of model-based clustering may provide better performance with less hassle.

Future research might explore additional ensemble configurations, including hybrid ensembles that include additional model-based and distance-based clustering procedures (e.g., PAM, DBSCAN, etc.), ensembles that vary the subsets of variables used as bases and ensembles for handling mixtures of metric and categorical variables (e.g., ensembles of model-based and Kamila solutions). As well, finding ways to incorporate real-world data sets would be most useful.

Keith Chrzan

Joseph White

REFERENCES

Chrzan, K., & White, J. (2021). "Replication of known segment structure and membership," Sawtooth Software Conference Proceedings, 217–226.

Chrzan, K., & White, J. (2022). "Variable selection in segmentation," Sawtooth Software Conference Proceedings, 207–218.

Morey, L., & Agresti, A. (1984). "The measurement of classification agreement: An adjustment to the Rand statistic for chance agreement," Educational and Psychological Measurement, 44, 33–37.

Orme, B., & Johnson, R. (2008). "Improving K-Means Cluster Analysis: Ensemble Analysis Instead of Highest Reproducibility Replicates," Sawtooth Software White Paper.

Retzer, J., & Shan, M. (2007). "Cluster ensemble analysis and graphical depiction of cluster partitions," Sawtooth Software Conference Proceedings, 239–250.

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). "mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models," The R Journal, 8(1), 289–317.

Strehl, A., & Ghosh, J. (2002). "Cluster ensembles—A knowledge reuse framework for combining multiple partitions," Journal of Machine Learning Research, 3, 583–617.

Vermunt, J. K., & Magidson, J. (2021). Latent GOLD Version 6.0 Upgrade Manual. Arlington, MA: Statistical Innovations Inc.

DETERMINING THE VALUE OF PRICE THRESHOLDS IN PRICING CONJOINT STUDIES

MICHAEL SMITH
JULI PHAM
SKIM

ABSTRACT

Clients across multiple industries have become intrigued by the idea of potential price thresholds or price cliffs in the price elasticities of their products and these thresholds can be very key to their pricing strategies. In this presentation, through testing the application of post-hoc and modelled threshold options, we will illustrate whether pricing thresholds add value to our pricing models, what type of pricing thresholds (post-hoc or modelled) work the best, and if there are types of pricing studies that pricing thresholds are most appropriate for.

INTRODUCTION

Over the past decade, numerous clients across diverse industries have become intrigued by the concept of potential price thresholds or price cliffs within the price elasticities of their products. The question arises: do their products have a specific price point at which demand drops significantly? These potential price thresholds or cliffs can be pivotal to our clients' pricing strategies.

One method to test price thresholds and cliffs within a conjoint pricing study is to apply a rule to the conjoint model utilities post-hoc by assigning a large negative value to the utilities for each respondent above the highest price of a product they selected. Another method is to include the highest price of a product they selected as a parameter within the conjoint model, allowing the model to determine the effect. While incorporating known information into the model may improve model fit, it could potentially lead to overfitting. It is essential to evaluate whether the constrained model (either post-hoc or added modelled parameter) results in a better model out-of-sample. Additionally, it is important to determine if different types of pricing studies benefit more or less from utilizing a price threshold or cliff.

We conducted tests across 8–10 studies to evaluate these hypotheses by comparing models with post-hoc thresholds and models with modeled thresholds. We assessed the RLH, holdout hit rates, and other model fit parameters to determine the value of applying thresholds within pricing studies and whether there is a difference between various types of pricing studies.

At the conclusion of our presentation, we aim to address the following questions:

1. Do pricing thresholds add value to our pricing models?
2. If pricing thresholds add value, which approach (post-hoc or modelled) is better?
3. Is the value of price thresholds consistent or do they work better for different types of pricing studies?

These findings will enable us to provide the most insightful pricing strategies to our clients.

DEFINITION AND HISTORY

We begin by defining price thresholds and price cliffs. A price threshold refers to a specific price point where consumers perceive a significant change in value. It is a key concept in psychological pricing, which aims to influence consumer perception and purchasing decisions. In essence, price thresholds are the point at which consumers might consider a product or service too expensive. A price cliff is defined as a substantial drop in demand at a certain price point. Clients often seek these discontinuities in demand as they provide clear guidance for pricing strategies.

Several common issues in pricing research have sparked our interest in the topic of price cliffs and price thresholds:

1. Unrealistic elasticities at high prices
2. High or overstated feature willingness to pay values
3. Big disparity in our simulated shares and in-market data
4. Continuing positive slopes to revenue curves at higher prices
5. No true revenue optimizing prices—no threshold point

Unrealistic elasticities at high prices and overstated willingness to pay are the most frequently encountered issues.

Price thresholds and price cliffs have been previously investigated. In 2007, Srinivasan, Pauwels, and Franses' paper "When Do Price Thresholds Matter in Retail Categories?" investigated price threshold elasticities for the top 4 brands in 20 different fast-selling consumer goods retail categories[1]. They determined that in 76% of the brand category combinations, that there was significance at either the historical or current price threshold and concluded that price thresholds do matter for a majority of brands and categories.

In 2023, Otter et al. presented at the Sawtooth Conference on the topic of budget constraints and whether traditional CBC models are biased by omitting budget constraints[2]. They developed a new Bayesian model that incorporates a latent effect for budget constraints in conjoint analysis for large-ticket items (such as automobiles and personal computers). They found empirical evidence that budgets are relevant and that omitting them from traditional CBC models leads to suboptimal and biased insights. Including budget constraints also improved the accuracy of pricing estimates and model predictions.

STUDY SETUP

In this research, we examine two different types of cliffing: Post-Hoc Cliffing and Cliffing in Estimation. For Post-Hoc Cliffing, the model is run as usual, but for prices exceeding the respondent's maximum price selected in the conjoint, we replace those utilities with a high negative number. For example, if respondent 1's maximum price selected was \$10, we replaced the utility of the \$20 price with -999. Cliffing in Estimation differs in that it accounts for the

---

[1] Pauwels, Koen & Srinivasan, Shuba & Franses, Philip. (2007). When Do Price Thresholds Matter in Retail Categories?. Marketing Science. 26. 83–100. 10.1287/mksc.1060.0207.

[2] Pachali, M. J., Kurz, P., & Otter, T. (2023). Omitted Budget Constraint Bias and Implications for Competitive Pricing. *Journal of Marketing Research*, *60*(5), 968–986. https://doi.org/10.1177/00222437221145283 (Original work published 2023)

maximum price a respondent selects and treats the penalty as additional data in estimation. This approach models with additional information, potentially altering variable effects on other non-price attributes. In execution, the choice data file includes an additional column that applies a penalty for options more expensive than the maximum price. For example, respondent 1 had a maximum price of $10, so in the choice data informing the estimation, we added a penalty column with 1's for all rows more expensive than $10 and 0's for all rows $10 or less.

We will explore these types of cliffing today and aim to answer the following questions for both types:

- Do price thresholds improve model performance? We will examine statistics such as Hit Rate, Mean Probability, Mean Average Error, Root Mean Squared Error, Log Likelihood, and Correlation to in-market data.
- How do price thresholds impact WTP values?
- How do price thresholds affect the importance of price?

We examined 12 different studies covering categories such as education tech, cloud storage, document management, oral care, skin care, and laundry care. Eight studies were multi-attribute studies, breaking product components down into more attributes (e.g., brand, size, flavor). In contrast, four studies were SKU-Price studies, where SKU is the core attribute encapsulating other features. In terms of price coding, eight studies coded prices as slopes, also known as unary or thermometer coding. Four studies used discrete price coding, where prices were coded as distinct separate levels. Most studies tested price ranges under a couple of hundred dollars, with about half testing price ranges below $50. Exceptions included Study C, which had a price range of $400 to $1400, and Study A, which displayed prices to respondents as a percentage of a base price rather than a monetary value.

ID	Category	Study Type	Price Type	# Price Levels	Price Range
Study A	Education Tech	Multi-Attribute	Discrete	5	-75% to 43%
Study B1	Cloud Storage	Multi-Attribute	Slope	6	$1.99 to 29.99
Study B2	Cloud Storage	Multi-Attribute	Discrete	7	$1.99 to 29.99
Study C1	Home Tech	Multi-Attribute	Slope	5	$399 to 1,399
Study C2	Home Tech	Multi-Attribute	Discrete	6	$399 to 1,399
Study D1	Cloud Storage	Multi-Attribute	Slope	7	$1.99 to 44.99
Study D2	Cloud Storage	Multi-Attribute	Discrete	6	$1.99 to 44.99
Study E	Document Management	Multi-Attribute	Slope	13	$4 to 40
Study F	Oral Care	SKU-Price	Slope	16	$3.39 to 16.09
Study G	Skin Care	SKU-Price	Slope	28	$5 to 340
Study H	Hair Care	SKU-Price	Slope	31	$34.95 to 159.95
Study I	Laundry Care	SKU-Price	Slope	8	$14.99 to 25.99

RESULTS

To explore and validate these two cliffing techniques, we looked at both in-sample and out-of-sample validation methods. In-sample testing evaluates model accuracy by holding out a few choice tasks per respondent, estimating the model on the remaining tasks, and comparing predictions for the holdouts to actual choices. Out-of-sample testing goes further by setting aside a portion of respondents entirely, estimating the model on the rest, and then evaluating how well the model predicts aggregate choices for the holdout tasks among the holdout group.

In-Sample Test Results

For in-sample validation, we evaluated six key metrics to assess the impact of cliffing on model performance: Hit Rate, Mean Probability, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Log-Likelihood, and percentage correlation with in-market product shares provided by clients.

Hit Rate

Hit rate evaluates the discrete choice accuracy of a model by measuring the proportion of choices in the holdouts that were correctly predicted by the conjoint model. An increase in hit rate suggests better model performance and prediction. Across studies, Cliffing in Estimation produced negligible differences in hit rate compared to standard models without any cliffing applied. In Table A, only two studies showed differences exceeding 1%, with most studies actually favoring the standard model on hit rate performance. However, again, the differences observed are generally insignificant.

Table A: Hit Rate (Cliffing in Estimation)

ID	Study Type	Price Type	# Price Levels	Change in Hit Rate
Study A	Multi-Attribute	Discrete	5	-0.365%
Study B1	Multi-Attribute	Slope	6	-0.199%
Study B2	Multi-Attribute	Discrete	7	0.223%
Study C1	Multi-Attribute	Slope	5	-0.590%
Study C2	Multi-Attribute	Discrete	6	-0.486%
Study D1	Multi-Attribute	Slope	7	-0.472%
Study D2	Multi-Attribute	Discrete	6	-0.770%
Study E	Multi-Attribute	Slope	13	-0.485%
Study F	SKU-Price	Slope	16	-0.411%
Study G	SKU-Price	Slope	28	0.794%
Study H	SKU-Price	Slope	31	-1.653%
Study I	SKU-Price	Slope	8	1.360%

Post-Hoc Cliffing showed similarly negligible effects, with differences emerging only at the second decimal place. The differences in Table B are very slight, suggesting that the model's choice accuracy after implementing Post-Hoc Cliffing remains comparable to standard models. Although very slight, these differences and the more consistent trend perhaps suggest Post-Hoc Cliffing performs marginally better than Cliffing in Estimation in choice accuracy.

Table B: Hit Rate (Post-Hoc Cliffing)

ID	Study Type	Price Type	# Price Levels	Change in Hit Rate
Study A	Multi-Attribute	Discrete	5	0.033%
Study B1	Multi-Attribute	Slope	6	0.298%
Study B2	Multi-Attribute	Discrete	7	0.670%
Study C1	Multi-Attribute	Slope	5	0.035%
Study C2	Multi-Attribute	Discrete	6	0.052%
Study D1	Multi-Attribute	Slope	7	0.025%
Study D2	Multi-Attribute	Discrete	6	0.025%
Study E	Multi-Attribute	Slope	13	0.000%
Study F	SKU-Price	Slope	16	0.565%
Study G	SKU-Price	Slope	28	-3.968%
Study H	SKU-Price	Slope	31	0.000%
Study I	SKU-Price	Slope	8	0.320%

Mean Probability

Mean probability reflects how confidently the model predicts actual choice and is used as a proxy for mean root-likelihood (mean RLH). An increase in mean probability is an indicator of improved model performance. As seen in Table C, results were mixed for Cliffing in Estimation.

Table C: Mean Probability (Cliffing in Estimation)

ID	Study Type	Price Type	# Price Levels	Change in Mean Prob.
Study A	Multi-Attribute	Discrete	5	-0.185%
Study B1	Multi-Attribute	Slope	6	0.567%
Study B2	Multi-Attribute	Discrete	7	1.082%
Study C1	Multi-Attribute	Slope	5	-0.248%
Study C2	Multi-Attribute	Discrete	6	-0.152%
Study D1	Multi-Attribute	Slope	7	-0.101%
Study D2	Multi-Attribute	Discrete	6	-0.197%
Study E	Multi-Attribute	Slope	13	0.216%
Study F	SKU-Price	Slope	16	2.169%
Study G	SKU-Price	Slope	28	0.434%
Study H	SKU-Price	Slope	31	-0.992%
Study I	SKU-Price	Slope	8	0.778%

Meanwhile in Table D, we see more consistent results that indicate Post-Hoc Cliffing may slightly improve mean probability compared to standard models. However, across both cliffing methods, the impact, positive or negative, was typically under 1–2%, reinforcing the minimal practical effect of either technique on predictive confidence.

Table D: Mean Probability (Cliffing in Estimation)

ID	Study Type	Price Type	# Price Levels	Change in Mean Prob.
Study A	Multi-Attribute	Discrete	5	0.454%
Study B1	Multi-Attribute	Slope	6	0.850%
Study B2	Multi-Attribute	Discrete	7	1.752%
Study C1	Multi-Attribute	Slope	5	0.159%
Study C2	Multi-Attribute	Discrete	6	0.279%
Study D1	Multi-Attribute	Slope	7	0.312%
Study D2	Multi-Attribute	Discrete	6	0.335%
Study E	Multi-Attribute	Slope	13	1.141%
Study F	SKU-Price	Slope	16	1.069%
Study G	SKU-Price	Slope	28	-2.234%
Study H	SKU-Price	Slope	31	0.163%
Study I	SKU-Price	Slope	8	0.748%

Prediction Error Metrics: MAE and RMSE

Mean absolute error (MAE) and root mean square error (RMSE) offer error-based assessments of prediction accuracy. A model with lower error, MAE or RMSE, is superior. Although not ideal at the respondent level due to limited data, these metrics remain useful as familiar, high-level benchmarks.

Across both metrics, Cliffing in Estimation led to mixed and minor effects as indicated by the third decimal place values in Table E and Table G. Post-Hoc Cliffing also produced small differences in magnitude for both error rates, but with more consistent directional improvements (Table F and Table H). More often, Post-Hoc Cliffing yielded lower MAE and RMSE than the standard model, especially in multi-attribute studies. This pattern suggests that Post-Hoc Cliffing may offer a modest advantage in reducing prediction error compared to Cliffing in Estimation. Notably, Study G again emerged as an outlier in both cliffing approaches, possibly due to the presence of slope groups.

Table E: Mean Absolute Error (Cliffing in Estimation)

ID	Study Type	Price Type	# Price Levels	Change in MAE
Study A	Multi-Attribute	Discrete	5	0.0015
Study B1	Multi-Attribute	Slope	6	-0.0032
Study B2	Multi-Attribute	Discrete	7	-0.0062
Study C1	Multi-Attribute	Slope	5	0.0025
Study C2	Multi-Attribute	Discrete	6	0.0015
Study D1	Multi-Attribute	Slope	7	0.0008
Study D2	Multi-Attribute	Discrete	6	0.0016
Study E	Multi-Attribute	Slope	13	-0.0008
Study F	SKU-Price	Slope	16	-0.0006
Study G	SKU-Price	Slope	28	-0.0002
Study H	SKU-Price	Slope	31	0.0012
Study I	SKU-Price	Slope	8	-0.0005

Table F: Mean Absolute Error (Post-Hoc Cliffing)

ID	Study Type	Price Type	# Price Levels	Change in MAE
Study A	Multi-Attribute	Discrete	5	-0.0024
Study B1	Multi-Attribute	Slope	6	-0.0049
Study B2	Multi-Attribute	Discrete	7	-0.0100
Study C1	Multi-Attribute	Slope	5	-0.0016
Study C2	Multi-Attribute	Discrete	6	-0.0028
Study D1	Multi-Attribute	Slope	7	-0.0025
Study D2	Multi-Attribute	Discrete	6	-0.0027
Study E	Multi-Attribute	Slope	13	-0.0042
Study F	SKU-Price	Slope	16	-0.0003
Study G	SKU-Price	Slope	28	0.0008
Study H	SKU-Price	Slope	31	-0.0002
Study I	SKU-Price	Slope	8	-0.0005

Table G: Root Mean Squared Error (Cliffing in Estimation)

ID	Study Type	Price Type	# Price Levels	Change in RMSE
Study A	Multi-Attribute	Discrete	5	-0.0026
Study B1	Multi-Attribute	Slope	6	-0.0023
Study B2	Multi-Attribute	Discrete	7	-0.0052
Study C1	Multi-Attribute	Slope	5	0.0007
Study C2	Multi-Attribute	Discrete	6	-0.0002
Study D1	Multi-Attribute	Slope	7	0.0013
Study D2	Multi-Attribute	Discrete	6	0.0011
Study E	Multi-Attribute	Slope	13	0.0023
Study F	SKU-Price	Slope	16	0.0009
Study G	SKU-Price	Slope	28	-0.0007
Study H	SKU-Price	Slope	31	0.0038
Study I	SKU-Price	Slope	8	-0.0010

Table H: Root Mean Squared Error (Post-Hoc Cliffing)

ID	Study Type	Price Type	# Price Levels	Change in RMSE
Study A	Multi-Attribute	Discrete	5	-0.0008
Study B1	Multi-Attribute	Slope	6	-0.0024
Study B2	Multi-Attribute	Discrete	7	-0.0048
Study C1	Multi-Attribute	Slope	5	-0.0005
Study C2	Multi-Attribute	Discrete	6	-0.0010
Study D1	Multi-Attribute	Slope	7	-0.0006
Study D2	Multi-Attribute	Discrete	6	-0.0007
Study E	Multi-Attribute	Slope	13	-0.0011
Study F	SKU-Price	Slope	16	-0.0005
Study G	SKU-Price	Slope	28	0.0056
Study H	SKU-Price	Slope	31	0.0001
Study I	SKU-Price	Slope	8	-0.0004

Log-Likelihood

Log-likelihood measures how well the model assigns probability to the observed choice. Because it is influenced by the number of parameters tested, which varies study to study, we assessed performance of this metric using percent change rather than absolute change to allow for more meaningful comparisons. A decrease in log-likelihood suggests better model performance.

As shown in Table I, Cliffing in Estimation led to minimal improvements over the standard model, with percentage changes in log-likelihood typically under 5% across all studies. A similar trend was observed for Post-Hoc Cliffing (Table J), which also showed limited differences overall. One exception was Study H, which emerged as a clear outlier in both magnitude and direction. This may be due to its design, which featured three holdout tasks—more than all other studies tested—resulting in less data to inform the model.

Table I: Log-Likelihood (Cliffing in Estimation)

ID	Study Type	Price Type	# Price Levels	% Change in LogL
Study A	Multi-Attribute	Discrete	5	-3.1%
Study B1	Multi-Attribute	Slope	6	-2.7%
Study B2	Multi-Attribute	Discrete	7	-6.7%
Study C1	Multi-Attribute	Slope	5	0.0%
Study C2	Multi-Attribute	Discrete	6	-0.4%
Study D1	Multi-Attribute	Slope	7	-0.4%
Study D2	Multi-Attribute	Discrete	6	-0.5%
Study E	Multi-Attribute	Slope	13	-0.1%
Study F	SKU-Price	Slope	16	23.9%
Study H	SKU-Price	Slope	31	-1.1%
Study I	SKU-Price	Slope	8	-3.2%

Table J: Log-Likelihood (Post-Hoc Cliffing)

ID	Study Type	Price Type	# Price Levels	% Change in LogL
Study A	Multi-Attribute	Discrete	5	-1.3%
Study B1	Multi-Attribute	Slope	6	-1.9%
Study B2	Multi-Attribute	Discrete	7	-3.3%
Study C1	Multi-Attribute	Slope	5	-0.4%
Study C2	Multi-Attribute	Discrete	6	-0.7%
Study D1	Multi-Attribute	Slope	7	-0.7%
Study D2	Multi-Attribute	Discrete	6	-0.8%
Study E	Multi-Attribute	Slope	13	-2.6%
Study F	SKU-Price	Slope	16	-2.1%
Study H	SKU-Price	Slope	31	81.2%
Study I	SKU-Price	Slope	8	-1.9%

Correlation with In-Market Data

The final metric evaluated for in-sample validation was the percentage correlation between predicted market shares from the conjoint model and actual in-market shares, using client-provided data. In-market data was only available for four of the studies in this research. Across all four, Cliffing in Estimation consistently reduced the correlation with real-world outcomes. In comparison, Post-Hoc Cliffing produced mixed results, with no clear pattern of improvement or decline, and overall showed no significant difference in correlation compared to standard models.

Table K: Correlation with In-Market Data

Cliffing	ID	Study Type	Price Type	# Price Levels	Change in Correlation
In Estimation	Study E	Multi-Attribute	Slope	13	-3.258%
In Estimation	Study G	SKU-Price	Slope	28	-37.718%
In Estimation	Study H	SKU-Price	Slope	31	-13.348%
In Estimation	Study I	SKU-Price	Slope	8	-0.461%
Post-Hoc	Study E	Multi-Attribute	Slope	13	0.008%
Post-Hoc	Study G	SKU-Price	Slope	28	-0.394%
Post-Hoc	Study H	SKU-Price	Slope	31	0.448%
Post-Hoc	Study I	SKU-Price	Slope	8	0.000%

Out-of-Sample Test Results

For out-of-sample validation, MAE was the primary metric used to assess the impact of cliffing techniques on model performance. Out-of-sample validation was possible in two studies in our research, as these were the only ones that included fixed tasks shown to all respondents. As shown in Table L, overall findings were mixed. Cliffing in Estimation produced larger differences in MAE, but these varied in direction across the studies. Post-Hoc Cliffing, by contrast, resulted in negligible differences at the fourth or fifth decimal place, suggesting that this technique neither meaningfully improved nor degraded model performance in these two test cases. Unfortunately, the small sample of studies limits our ability to draw definitive conclusions or identify trends on these cliffing approaches.

Table L: Mean Absolute Error (Out-of-Sample)

Cliffing	ID	Study Type	Price Type	# Price Levels	Change in MAE
In Estimation	Study E	Multi-Attribute	Slope	13	0.01306
In Estimation	Study H	SKU-Price	Slope	31	-0.00448
Post-Hoc	Study E	Multi-Attribute	Slope	13	-0.00003
Post-Hoc	Study H	SKU-Price	Slope	31	0.00011

Practical Outcomes of Price Thresholds

Having established how cliffing techniques affect model performance, we now turn to their practical outcomes, specifically, how they influence willingness to pay and the importance of price. While model accuracy is paramount to researchers, analyses such as willingness to pay and attribute importance tend to be more meaningful to clients and have practical implications, as they shape how customers are understood and directly inform pricing strategies and product positioning.

Willingness to Pay (WTP)

Willingness to pay (WTP) is the maximum price a consumer is willing to pay for a product or feature. To conduct this analysis, we leveraged Sawtooth's Choice Simulator and Sampling of Scenarios method to calculate the WTP of each feature, with and without cliffing applied to the model. Then, for each feature, we calculate the percentage change in willingness to pay when we cliff versus the standard model. Lastly, we average the percentage change in WTP across all features. Since we used Sawtooth's Choice Simulator to calculate WTP, we only evaluated multi-attribute studies where prices were coded as discrete.

Our results showed that Cliffing in Estimation led to inconsistent changes in willingness to pay values. This may be due to the added constraints and additional model information, such as maximum price inputs and penalty data, which influences the coefficients of non-price attributes as well. This results in greater variability in the share impact of features, thus influencing the WTP. In contrast, Post-Hoc Cliffing lowered WTP values, though the magnitude varied. One study (Study A) saw a more substantial WTP drop, possibly because price in this study was expressed as a percentage of a current price rather than as an absolute monetary value.

Table M: Willingness to Pay (WTP)

Cliffing	ID	Study Type	Price Type	# Price Levels	Avg % Change in WTP
In Estimation	Study A	Multi-Attribute	Discrete	5	9.6%
In Estimation	Study B2	Multi-Attribute	Discrete	7	1.7%
In Estimation	Study C2	Multi-Attribute	Discrete	6	-40.0%
In Estimation	Study D2	Multi-Attribute	Discrete	6	4.6%
Post-Hoc	Study A	Multi-Attribute	Discrete	5	-25.8%
Post-Hoc	Study B2	Multi-Attribute	Discrete	7	-0.7%
Post-Hoc	Study C2	Multi-Attribute	Discrete	6	-0.2%
Post-Hoc	Study D2	Multi-Attribute	Discrete	6	0.0%

Price Importance

Price importance reflects the impact of price on preference share relative to other attributes tested in the conjoint. It is calculated by taking the difference in preference share between the most and least preferred levels of each attribute, then converting those differences into percentages to express their relative impact. This approach allows us to compare the importance of price against other attributes on a standard scale.

Cliffing in Estimation produced mixed results for price importance (Table N). Because this cliffing technique modifies the underlying data used to model, it can affect the utilities of both price and non-price attributes, leading to unpredictable impact to preference share. Post-Hoc Cliffing, on the other hand, consistently increased the relative importance of price (Table O). This outcome is expected, as the post-hoc method effectively flattens utilities at certain price levels, which can only shift the model in one direction—making price appear more influential. However, the increase of price importance is modest and was typically less than 5%. Study A once again stands out as an outlier.

Table N: Price Importance (Cliffing in Estimation)

ID	Study Type	Price Type	# Price Levels	% Change in Price Importance
Study A	Multi-Attribute	Discrete	5	38.793%
Study B1	Multi-Attribute	Slope	6	-19.708%
Study B2	Multi-Attribute	Discrete	7	-1.157%
Study C1	Multi-Attribute	Slope	5	-0.847%
Study C2	Multi-Attribute	Discrete	6	-0.456%
Study D1	Multi-Attribute	Slope	7	1.007%
Study D2	Multi-Attribute	Discrete	6	0.323%
Study E	Multi-Attribute	Slope	13	-3.198%

Table O: Price Importance (Post-Hoc Cliffing)

ID	Study Type	Price Type	# Price Levels	% Change in Price Importance
Study A	Multi-Attribute	Discrete	5	42.180%
Study B1	Multi-Attribute	Slope	6	0.897%
Study B2	Multi-Attribute	Discrete	7	4.121%
Study C1	Multi-Attribute	Slope	5	1.421%
Study C2	Multi-Attribute	Discrete	6	1.719%
Study D1	Multi-Attribute	Slope	7	0.187%
Study D2	Multi-Attribute	Discrete	6	0.084%
Study E	Multi-Attribute	Slope	13	3.555%

Client Reactions

Anecdotally, our clients have responded positively to the inclusion of price cliffing in our models, specifically Post-Hoc Cliffing. Clients find the resulting discontinuities in demand curves insightful, especially when revenue-maximizing prices fall below the highest tested price points, which can help build confidence in pricing recommendations. That said, the effectiveness of this thresholding approach depends heavily on testing a sufficient range of price points. If cliffs exist at untested price levels, the method will fail to capture them. As such, when considering price thresholding in a study, it's critical to carefully select price points.

CONCLUSION

In summary, our findings suggest that incorporating price thresholds does not materially improve model performance, with both Cliffing in Estimation and Post-Hoc Cliffing showing minimal and inconsistent effects on standard accuracy metrics. Among the two methods, Post-Hoc Cliffing slightly outperformed Cliffing in Estimation and more reliably dampened willingness to pay and increased price importance. This makes the technique more helpful when outputs seem inflated or unintuitive. While results varied across studies, Post-Hoc Cliffing is a low-risk, model modification that can enhance interpretability for clients. Its effectiveness depends on thoughtful price selection and time for iteration, making it a useful, but not universally necessary, addition to the modeling toolkit.

Michael Smith

Juli Pham

LEVERAGING THE 4 P MARKETING FRAMEWORK TO CALIBRATE CONJOINT MODELS

JAMES PITCHER
DIMITRI LIAKHOVITSKI
ALEXANDRA CHIRILOV
NIELSEN IQ

ABSTRACT

Conjoint analysis is a powerful tool for modelling consumer preferences, yet its predictions often diverge from real-world market shares due to unaccounted in-market factors. While calibration using distribution (Place) is common, this paper explores whether incorporating all four elements of the 4 P marketing framework—Product, Price, Place, and Promotion—can further improve the alignment between conjoint-derived brand preference shares and actual market shares. Using Nielsen IQ's global Point-Of-Sale data and survey-based conjoint studies across two product categories and five countries, we tested multiple calibration approaches, including utility scaling, direct adjustment using 4 P metrics, and a regression-based method. Results show that while Place and utility scaling significantly reduce the discrepancy between conjoint based preference shares and market shares, additional calibration using Product, Price, and Promotion can further improve accuracy in some cases. Although measures of the 4Ps are best derived from sales data, survey-derived 4 P metrics also show promise, when sales data is unavailable. We conclude that the 4 P framework offers a scalable and theoretically grounded path to enhancing conjoint model calibration, though further research is needed to refine the calibration methodology.

MOTIVATION

Conjoint analysis is widely used in market research to model consumer decision-making and predict market outcomes. Despite its robustness, one common challenge in conjoint analysis is the divergence between the shares of preference generated by the model and the observed market shares.

Traditionally, it has been assumed that the difference between conjoint shares and real-world market shares is largely because in conjoint analysis awareness and distribution of brands and products is 100%. Whereas in reality, brands and products are not available all of the time in every store and consumers are not aware of all brands and products that exist on the market. Hence, conjoint shares are commonly calibrated using awareness and distribution in the hope that the resulting calibrated shares of preference will come closer to real-world market shares.

In our 2021 Sawtooth Conference paper (*Chirilov and Pitcher, 2021*), we explored the impact of calibrating conjoint shares by awareness and distribution. We found that adjusting conjoint shares of preference by distribution brought them closer to market shares. However, we found that calibrating by awareness does not bring conjoint preferences closer to market shares, because conjoint already captures the role of awareness in the choices respondents make. Furthermore, in FMCG, calibrating by awareness can increase the error versus market shares because consumers will sometimes buy products that they are not aware of.

Although calibrating by distribution brings conjoint shares of preference closer to market shares, there are other reasons, beyond just distribution, to why we see differences between conjoint shares and market shares. In-market factors alter preferences at the point of purchase causing differences between conjoint preferences and market shares.

The well-established "4 P" marketing framework provides a comprehensive view of the main in-market factors that influence consumer choices. The 4 Ps are as follows:

- **Product:** the extent to which a brand offers products that meet consumer needs
- **Price:** A brand's price and price discounts
- **Place:** The ease of finding and purchasing brand's products offline or online
- **Promotion:** The effectiveness of how a brand communicates its offerings through various media channels

Calibrating by distribution accounts for Place only. But why should we calibrate by the other 3 Ps as well?

Let's start with Product. Although we can test different products in a conjoint study, it is impossible to simulate the entire market, even in an SKU-price conjoint. So, you end up misrepresenting the actual shelf presence in conjoint tasks. For example, maybe 50% of the SKUs on the market are Samsung but we do not show this in the conjoint exercise. Also, the overall reputation of *all* of a brand's products rubs off on products that are shown in conjoint exercise.

Next, let's talk about why we should calibrate by Price. Although we can test the impact of price in conjoint analysis, it is more difficult to test the impact of price promotions. Yes, we can test price promotions as an attribute in a conjoint study, but conjoint is a snapshot at one single moment in time for a limited number of SKUs. We cannot fully capture the market conditions, i.e., the level to which products were discounted across the year and across different stores. We also cannot test every product in conjoint studies and all types of price discounting.

And finally, let's talk about Promotion. Short-term advertising can influence what brand or product we buy. For example, if I've seen an advert encouraging me to buy an LG product recently, I'm more likely to buy LG. This advertising is not long-lasting, and its influence is not captured in the conjoint exercise. Choices in a conjoint study are only influenced by long-held feelings towards brands and products.

Given the influence of all 4 Ps on sales, we therefore designed a study to see if we can improve the calibration of conjoint shares by including elements of all 4 Ps. That is, not just calibrating our conjoint models by distribution (Place) alone, but also calibrating by measures of Product, Price and Promotion as well.

METHODOLOGY

At Nielsen IQ, we collect Point-of-Sales (POS) data in multiple categories of products in multiple countries around the world. We leveraged the POS data for this study focusing on one FMCG category (Chocolate Bars) and one Tech and Durables category (Large Screen TVs).

POS data were used to derive brand level KPIs for three of the four Ps: Product, Place, and Price (discount). Each major brand got a single score for each of the 4 Ps:

- *Product* KPI was operationalized as a combination of four aspects of each brand's product portfolio:
  - *Portfolio size*: the overall number of products in the brand's portfolio
  - *Diversity*: the range and diversity of products in the portfolio: how well it manages to address different consumer needs
  - *Focus*: the portfolio's concentration on specific consumer needs, demonstrating leadership in certain groupings of needs
  - *Innovation/Agility*: the portfolio's alignment with current market trends and the speed of introducing new products and features
- *Place* was operationalized as weighted distribution, i.e., % of stores the brand sells in, weighted by the total sales of each store.
- *Price (Discount)*: was operationalized as the multiplication of the average discount % and discount frequency.

The fourth P (*Promotion*) could not be derived from the POS data and thus was measured in consumer surveys. For each brand respondents were aware of, they were asked, if they recall seeing the brand mentioned in various media channels, such as in an online ad, on social media, in an online product review etc.

Table 1 below shows the two product categories, the five countries, the number of brands per cell (country/category), and for which of the 4 Ps we collected brand KPIs.

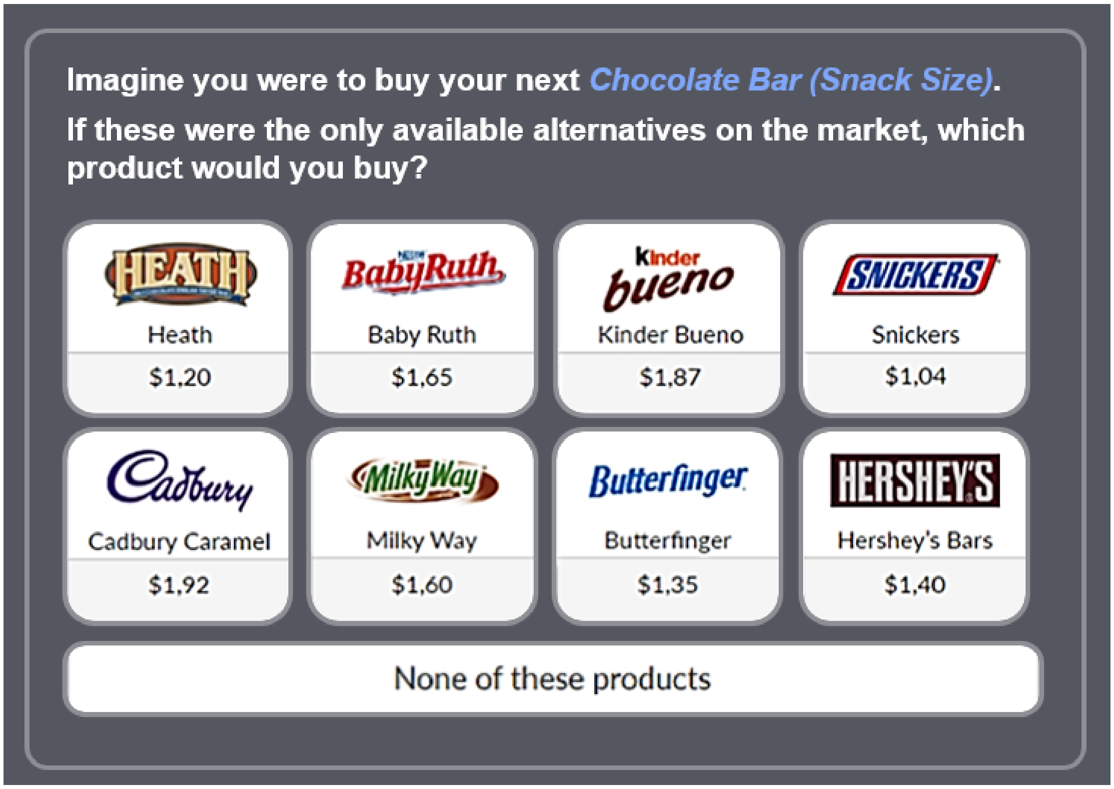
Table 1: Five cells of the current study and information collected for each cell.

Category	Country	No. of brands	Product	Place	Price	Promotion
Chocolate Bars	US	14	✓	✓	✓	✓
TV	Thailand	12	✓	✓	✓	✓
TV	France	15	✓	✓	✓	✗
TV	Italy	15	✓	✓	✓	✗
TV	UK	9	✓	✓	✓	✗

Conjoint Setup in the Survey

For each study cell (country/category combination) mentioned above we fielded a survey with category purchasers. Survey respondents went through a simple choice-based brand-price conjoint exercise where they were asked to imagine they were looking to buy a typical product in the category and then had to choose the product of which brands they would purchase (on several screens), for example, see Figure 2.

Figure 2: A screenshot of one task of the Brand-Price Conjoint exercise.



Prices were conditional (on brand) with 5 levels (from 20% below the brand's mid-price, i.e., its current market price, to 20% above the mid-price).

Hierarchical Bayesian estimation—in Choice Model R—was used to derive utilities for Brands and Price. The utilities converged nicely. For each study cell a "base case" scenario was created in which each brand had its current market price and the shares of preference from that scenario were then used as a point of comparison to try to match the actual market shares of the brands.

Results

4 P Calibration by Adjusting Respondent Level Utilities and Shares

Prior research has demonstrated that preference shares could be calibrated to better match market shares using a utility scaling factor and distribution (*Place*). That's what we've also done, as follows:

1. Multiply each respondent's base case scenario's probability of selecting the brand by that brand's weighted distribution on the market:

Figure 3: Adjusting respondents' choice probabilities by brand's distribution.

Respondent	SAMSUNG	ACONATIC	TCL	LG	HISENSE	...
1	3.4%	29.5%	0.8%	0.1%	0.1%	...
2	19.5%	0.3%	0.7%	0.9%	3.0%	...
3	83.3%	0.5%	0.1%	0.0%	0.0%	...
4	5.1%	1.4%	1.0%	0.2%	0.0%	...
5	47.2%	0.8%	15.0%	6.4%	0.8%	...
...	...	...	...	...	...	...

	X	X	X	X	X	
Distribution:	95%	39%	47%	83%	73%	...
	=	=	=	=	=	

Respondent	SAMSUNG	ACONATIC	TCL	LG	HISENSE	...
1	3.2%	11.5%	0.4%	0.1%	0.1%	...
2	18.5%	0.1%	0.3%	0.7%	2.2%	...
3	79.1%	0.2%	0.0%	0.0%	0.0%	...
4	4.8%	0.5%	0.5%	0.2%	0.0%	...
5	44.8%	0.3%	7.1%	5.3%	0.6%	...
...	...	...	...	...	...	...

2. Rescale each respondent's resulting probabilities of choosing the brand so that they sum up to 100% again:

Figure 4: Rescaling individual choice probabilities to sum up to 100%.

Respondent	SAMSUNG	ACONATIC	TCL	LG	HISENSE	...	Sum
1	3.2%	11.5%	0.4%	0.1%	0.1%	...	**66.7%**
2	18.5%	0.1%	0.3%	0.7%	2.2%	...	**41.9%**
3	79.1%	0.2%	0.0%	0.0%	0.0%	...	**92.8%**
4	4.8%	0.5%	0.5%	0.2%	0.0%	...	**55.0%**
5	44.8%	0.3%	7.1%	5.3%	0.6%	...	**73.1%**
...	...	...	...	...	...	...	...

Rescale to 100%

Respondent	SAMSUNG	ACONATIC	TCL	LG	HISENSE	...	Sum
1	4.9%	17.3%	0.6%	0.1%	0.1%	...	**100.0%**
2	43.8%	0.3%	0.8%	1.8%	5.1%	...	**100.0%**
3	85.2%	0.2%	0.0%	0.0%	0.0%	...	**100.0%**
4	8.8%	1.0%	0.9%	0.3%	0.0%	...	**100.0%**
5	61.3%	0.4%	9.7%	7.3%	0.8%	...	**100.0%**
...	...	...	...	...	...	...	...

3. Then, we derived the best "utility scaling factor," i.e., a single multiplier that the original utilities were multiplied by, before the conjoint shares of preferences for the base case scenario were calculated (and then adjusted based on brands' distribution—as described above). For that, Excel's Solver was used with an objective to minimize the mean absolute deviation between the base case scenario's preference shares (adjusted based on the utility scaling factor and brands' distribution) and the actual market shares. As a result of Step 3 we had preference shares adjusted based on the *optimal* utility scaling factor and brands' distribution.

The optimal utility scaling factor varied across study cells between 0.26 (the lowest) and 0.82 (the highest).

As expected, in each cell of our study the usage of the optimal utility scaling factor and a further calibration based on Place (Distribution) helped move conjoint preference shares closer to the market shares—as demonstrated by the reduction of MAE (Mean Absolute Error)—see Table 2.

Table 2: Mean Absolute Error (MAE) of uncalibrated conjoint preference shares and preference shares calibrated based on the optimal scaling factor and distribution.

Mean Absolute Error (MAE) for each model

Cell	Uncalibrated	Distribution + Scale Factor
US Chocolate Bars	2.4%	2.3%
Thailand TV	6.0%	4.6%
France TV	3.9%	2.3%
Italy TV	3.5%	2.0%
UK TV	5.6%	4.5%

4. The shares from Step 3 were then *further* calibrated using brands' standings on the remaining 3 Ps (Product, Price (Discount) and Promotion—if available) of each brand (in a fashion similar to steps 1–2 above).

Table 3 below shows the MAE after further calibration using each of the remaining 3 Ps one by one (columns 3 through 6), using both Price (Discount) and Promotion (column 7) together, and using all 3 Ps together. Light green is used to highlight MAEs that are smaller than those in column 3.

Unfortunately, quite a few cells in the table below are *not* green: in most cases, calibrating using Place (Distribution) and the optimal scaling factor alone was sufficient. Further calibration using the remaining 3 Ps had little positive impact.

Table 3: Table 2 augmented with MAEs from calibrations based on additional Ps.

Mean Absolute Error (MAE) for each model

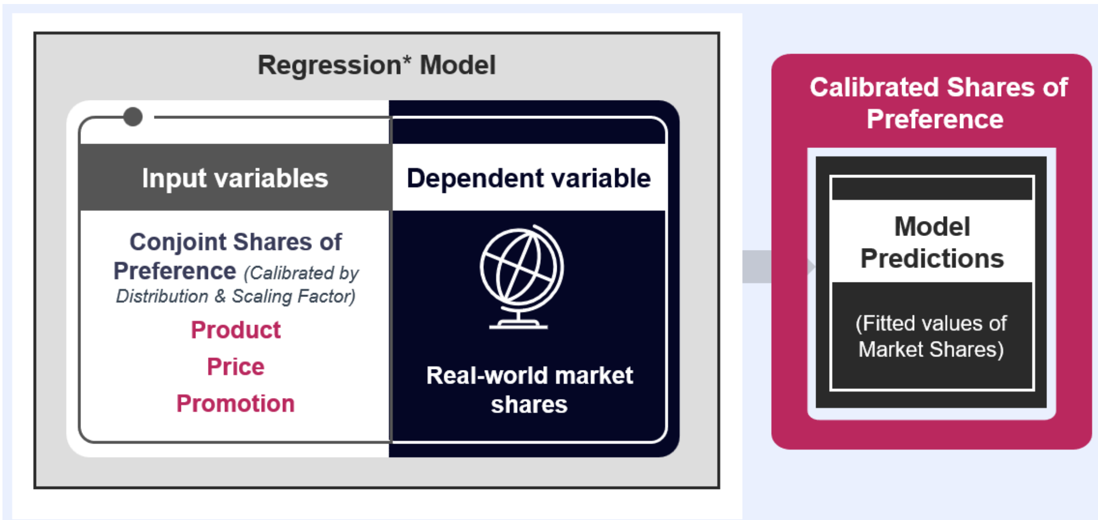
Cell	Uncalibrated	Distribution + Scale Factor	+ Price	+ Product	+ Promotion	+ Price + Product	+ Price + Product + Promotion
US Chocolate Bars	2.4%	2.3%	2.9%	**1.9%**	2.3%	**1.8%**	**1.9%**
Thailand TV	6.0%	4.6%	**4.3%**	6.8%	5.7%	5.7%	5.7%
France TV	3.9%	2.3%	3.0%	4.0%	NA*	3.4%	NA*
Italy TV	3.5%	2.0%	2.6%	4.4%	NA*	3.1%	NA*
UK TV	5.6%	4.5%	**4.1%**	6.4%	NA*	4.8%	NA*

While this result seemed disappointing, we felt that maybe the method of preference share calibration based on 3Ps that we used was not the optimal one. That's why we made another attempt to check if information contained by 3Ps could be used to move the conjoint preference shares closer to the market shares.

4 P Calibration Using Regression Analysis at the Aggregate (Brand) Level

In each study cell, we've run the following regressions (see Figure 5 and the following description)

Figure 5: Using regression analysis at the aggregate level to predict real market shares of brands based on calibrated conjoint market shares.



Staying at the level of individual brands (and thus, dealing with small sample sizes with 9 to 15 brands per cell), we regressed the actual market shares onto the following predictors:

- Predictor 1: Brands' shares of preference calibrated using the optimal utility scaling factor and distribution
- Adding to Predictor 1 one additional predictor: Brands' KPIs on Price (Discount), or Product, or Promotion (one at a time)
- Adding to Predictor 1 Brands' KPIs on two additional predictors: Price (Discount) and Product
- Adding to Predictor 1 Brands' KPIs on three additional predictors: Price (Discount), Product, and Promotion

In order to avoid overfitting, we used hold-one-out cross validation (see Figure 6 below): the regression model was developed always holding one brand out and then using the resulting regression model to predict the market share for the brand that was held out. Then we calculated MAE across all such regression runs (as many runs as the number of brands in the study cell).

Figure 6: Hold-One-Out Cross Validation

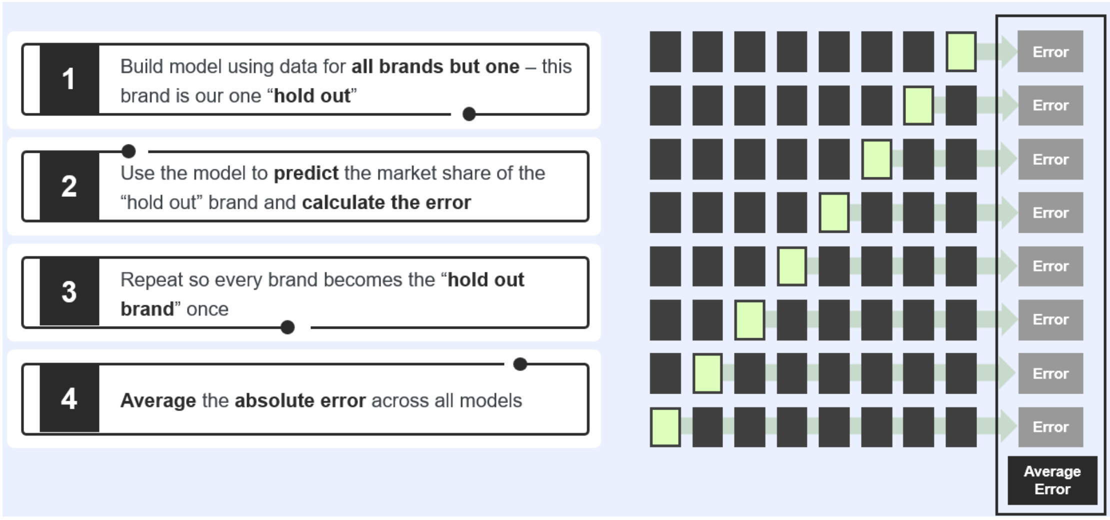


Table 4 below shows the resulting mean absolute errors. As we can see, for Thailand and France TVs, adding Price to the regression further reduce the errors. For Italy TVs, adding Product also helped. For US Chocolate Bars, adding Promotion helped reduce the error. And for all TV cells, adding both Price and Product to the regression equation helped reduce the error.

Table 4: MAEs of Different Regression Models

Mean Absolute Error (MAE) for each model

Cell	Distribution + Scale Factor	+ Price	+ Product	+ Promotion	+ Price + Product	+ Price + Product, + Promotion
US Chocolate Bars	5.8%	9.1%	9.2%	**2.8%**	10.8%	10.5%
Thailand TV	5.1%	**2.4%**	5.9%	6.5%	**3.5%**	**4.4%**
France TV	4.8%	**4.7%**	5.7%	NA*	**4.7%**	NA*
Italy TV	3.6%	4.0%	**1.7%**	NA*	**2.4%**	NA*
UK TV	5.8%	4.7%	9.2%	NA*	**5.1%**	NA*

These results are encouraging because they indicate that the information contained in the 3Ps that are not Place (Distribution) can help move conjoint preference shares closer to the actual market shares.

However, we are not arguing that using regressions the way we used them is the appropriate method of preference share calibrations. In their 2006 paper Johnson and Orme showed that post-hoc aggregate share adjustments can have undesired effects: they can change the fundamental relationships among products in the scenario in terms of self-elasticity and cross-elasticity. In addition, in some cases, such adjustments can lead to an increase in price of an SKU not only reducing its own share but also reducing the share of a different SKU.

We hope that researchers could continue our research to help find a more appropriate way of leveraging 4Ps to move conjoint shares closer to market shares.

4 P Calibration Using Survey Data Only

Figure **7** shows the correlation between the measures of all 4 Ps derived from survey and sales data. The survey measures of Place and Product align well with the measures derived from sales data. For example, the survey measure of "Place" for US Chocolate bars has a correlation of 0.9 with the weighted distribution of the brands tested. However, the survey measure of Price aligns less well, especially for US Chocolate Bars.

Figure 7: Correlation between metrics derived from survey and from sales data.

	Product	Price	Place	Promotion
US Chocolate Bars	0.88	0.13	0.90	N/A (Promotion was always a survey-derived measure)
Thailand TV	0.73	0.44	0.66	

Figures 8 and 9 below show the impact of calibrating the conjoint shares using survey measures of the 4 Ps. In Figure 8, we again assume that we have first calibrated the conjoint shares by Distribution and Scale Factor. We then additionally calibrate by the other 3Ps from survey data. For US Chocolate Bars, calibrating by Promotion, in addition to Distribution and Scale factor, yields the lowest MAE, of 2.8%. For Thailand TVs, calibrating by all 4 Ps gives the lowest MAE of 3.1%. Therefore, calibrating by more Ps does seem to generally improve the accuracy of the conjoint shares.

Figures 9 shows the errors for the scenario where we don't know the distribution of brands and we need to measure Place through the survey as well. For US Chocolate Bars, once we have calibrated by the Scale Factor, calibrating by the 4 Ps does not improve the accuracy any further. However, for Thailand TVs, calibrating by additional Ps does improve the accuracy further. Calibrating by all 4 Ps gives the lowest MAE of 4%.

Figure 8: MAE of the conjoint models calibrated by Distribution and Scale Factor and then by the other 3 Ps from survey data.

Mean Absolute Error (MAE) for each model

Cell	Distribution + Scale Factor	+ Price	+ Product	+ Promotion	+ Price + Product	+ Price + Product + Promotion
US Chocolate Bars	5.8%	6.4%	5.1%	2.8%	4.6%	3.2%
Thailand TV	5.1%	5.8%	5.8%	6.5%	3.2%	3.1%

Figure 9: MAE of the conjoint models calibrated by Scale Factor only and then by all 4 Ps from survey data.

Mean Absolute Error (MAE) for each model

Cell	Scale Factor only	+ Place	+ Price	+ Product	+ Promotion	+ Place + Price	+ Place + Price + Product	+ Place + Price + Product + Promotion
US Chocolate Bars	3.6%	4.0%	4.0%	3.9%	4.0%	4.2%	3.6%	4.8%
Thailand TV	7.4%	6.6%	**4.6%**	8.3%	7.9%	**5.3%**	**4.3%**	**4.0%**

To summarize:

- Survey Place and Product align well with the measures derived from sales data, although Price aligns less well.
- Price discounts are temporary and thus are, perhaps, more difficult for respondents to accurately recollect.
- Calibrating conjoint shares using the 4 P metrics derived from simple survey questions can improve their accuracy, although, as mentioned earlier, we still need to find the best method on how to incorporate them into the calibration process.

Future Research

Further work is required to develop a better method of calibrating conjoint-based shares of preference that incorporates all 4 Ps. We've already mentioned why we should avoid post-hoc aggregate share adjustments. So, the question remains, how can we best combine the 4 Ps into a single adjustment factor than can be applied at the respondent level?

We have simply taken an average of the scores and multiplied the conjoint shares by them. Do the 4 P scores need rescaling in some way? Should some Ps carry more weight than others? Is there a better method than simple multiplication with shares?

In addition, can we further improve the way we measure the 4 Ps using survey metrics? Can we find survey measures of the 4 Ps that align even better with metrics from sales data? In particular, can we find a better way to measure Price Discounts? As the survey measure of Price aligned least with the measure of Price from sales data.

Conclusions

Generally, calibrating by just distribution and the Scale Factor is sufficient to substantially improve the accuracy of conjoint shares of preference in relation to real-world market shares. However, we have seen that there is evidence that calibrating by all 4 Ps can improve the accuracy of conjoint shares further beyond calibrating by just distribution and the Scale Factor. But we haven't yet found the best method on how best to incorporate all 4 Ps into the calibration process. It is unlikely that there is a single solution that fits every market, and future research may also focus on better distinguishing the dynamics of different markets.

The fact a few simple additional survey questions can improve the accuracy of our conjoint models is potentially exciting news. The 4 Ps are best measured using sales data, but not everyone will always have access to sales data to be able to calculate the 4 P metrics. So, the fact we can derive them from a survey, makes the solution much more scalable.

In summary, we don't yet have all the answers on how best to calibrate conjoint models using the 4 Ps, but we believe the 4 Ps framework demonstrates exciting potential on how we can further improve our conjoint models.

James Pitcher

Dimitri Liakhovitski

Alexandra Chirilov

References

Chirilov, A & Pitcher, J. (2021). *Upgrade your Brand Tracker using the Power of Conjoint Analysis*, Sawtooth Software Conference 2021, San Antonio, Texas.

Johnson, R. & Orme, B. (2006). *External Effect Adjustments in Conjoint Analysis*, https://sawtoothsoftware.com/resources/technical-papers/external-effect-adjustments-in-conjoint-analysis