

Unleashing Innovation in Market Research

Using synthetic data to solve client problems

Julia Brannigan

Kerry Jones

ESOMAR

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@esomar.org

Website: www.esomar.org

Publication Date: September 2024

ESOMAR Publication Series Volume S416 Congress 2024

Copyright

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system of any nature, or transmitted or made available in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission of ESOMAR. ESOMAR will pursue copyright infringements.

In spite of careful preparation and editing, this publication may contain errors and imperfections. Authors, editors and ESOMAR do not accept any responsibility for the consequences that may arise as a result thereof.

The views expressed by the authors in this publication do not necessarily represent the views of ESOMAR.

By the mere offering of any material to ESOMAR in order to be published, the author thereby guarantees:

- that the author - in line with the ICC/ESOMAR International Code of Marketing and Social Research- has obtained permission from clients and/ or third parties to present and publish the information contained in the material offered to ESOMAR;
- that the material offered to ESOMAR does not infringe on any right of any third party; and
- that the author shall defend ESOMAR and hold ESOMAR harmless from any claim of any third party based upon the publication by ESOMAR of the offered material.

Published by ESOMAR, Amsterdam,

The Netherlands

Edited by: Bel Kerkhoff-Parnell, PhD

About ESOMAR

ESOMAR is the champion of the insights and analytics sector. It is the business community for every sector professional. Founded in 1947, the global membership association is a network reaching over 40,000 professionals and 750+ companies in 130+ countries. We support individual and corporate members supplying or using insights by helping them raise ethical standards, grow the demand for insights, and improve its uses and applications by all decision-makers.

About ESOMAR Membership

ESOMAR is open to everyone, all over the world, who believes that high quality research improves the way businesses make decisions. Our members are active in a wide range of industries and come from a variety of professional backgrounds, including research, marketing, advertising and media.

Membership benefits include the right to be listed in the ESOMAR Directories of Research Organisations and to use the ESOMAR Membership mark, plus access to a range of publications (either free of charge or with discount) and registration to all standard events, including the Annual Congress, at preferential Members' rates.

Members have the opportunity to attend and speak at conferences or take part in workshops. At all events the emphasis is on exchanging ideas, learning about latest developments and best practice and networking with other professionals in marketing, advertising and research. CONGRESS is our flagship event, attracting over 1,000 people, with a full programme of original papers and keynote speakers, plus a highly successful trade exhibition. Full details on latest membership are available online at www.esomar.org.

Contact us

ESOMAR

ESOMAR Office:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

The Netherlands

Tel.: +31 20 664 21 41

Email: info@esomar.org

Website: www.esomar.org

Unleashing Innovation in Market Research

Using synthetic data to solve client problems

Julia Brannigan

Kerry Jones

Introduction

The rapid digitisation of data in recent decades has bestowed upon us a treasure trove of information. However, the escalating need for comprehensive insights has birthed a novel approach: synthetic data. This revolutionary concept, leveraging computer-generated information, redefines the boundaries of market research innovation. At its core, synthetic data represents an innovative methodology that augments conventional market research strategies. By generating data that mimics real-world scenarios, it allows researchers to explore diverse consumer behaviours, preferences, and trends without compromising sensitive information or facing ethical constraints. Our exploration highlights the transformative potential of synthetic data in helping clients navigate confidentiality concerns, whilst amplifying the speed of iteration and enhancing the depth of analysis.

We have run regular hackathon sessions to stress test synthetic data against real-time primary data. We have tracked and monitored how our clients are using synthetic data off the back of these sessions. This paper explores the application of synthetic data in the context of a survey conducted for British food retailer, Marks and Spencer (M&S). Our focus is on comparing real-world data amongst M&S sushi buyers and non-buyers, in order to evaluate the efficacy and reliability of synthetic sampling in replicating the nuances of a specific consumer segment. We emphasise the ethical considerations and cautionary measures needed in harnessing the potential of synthetic data. It's not merely about leveraging technology, but ensuring that innovation aligns with ethical standards, upholds empirical validity and respects the dynamism of real-world scenarios.

M&S sushi case study: The context and challenge

Greencore supplies various food-to-go items to British retailer Marks & Spencer (commonly known as M&S), including pre-packed sushi. M&S is renowned for its premium food offerings and has seen a growing interest in its sushi products. Understanding attitudes to sushi, motivations and barriers to purchase and views on the sushi offer at M&S is critical for the company to refine its marketing, promotional and merchandising strategies, as well as develop the optimum product offer within sushi. However, acquiring a robust sample of M&S sushi buyers presents significant challenges due to their relatively niche status within the broader customer base. Further, M&S food buyers generally have a different profile to regular supermarket shoppers, e.g., they are an older demographic and less likely to do a main, big grocery shop at M&S.

As such, it was decided that it would be relevant and interesting to run a mirror survey to both real-life consumers and to a synthetic sample to see how they compare. If the results from the synthetic sample were favourable, then M&S would find it valuable to have robust numbers from the synthetic sample and could use it for highly confidential new product development. The primary hurdle in our study was the limited availability of M&S sushi buyers willing to participate in surveys. Traditional data collection methods often fall short in yielding sufficient numbers, leading to potential biases and limitations of insights derived. This scarcity necessitated an innovative approach to ensure comprehensive and actionable market research.

Building the synthetic sample

To address the sample size constraint, we turned to synthetic data generation. Synthetic data, created through advanced algorithms, aims to mirror the characteristics and behaviours of real-world consumers without requiring large-scale, real-world participation. The sample was based on eight Greencore consumer segments identified by attitudes to food, incorporating different views on meal preparation and planning, experimentation and sustainability and five sushi typologies (virgins, functionalists, pragmatists, experience seekers and purists). By generating a synthetic sample, we sought to augment our dataset, providing a more robust foundation for analysis.

Synthetic data engine overview

Our engine for generating synthetic survey responses is designed to gain consumer insights without the need for extensive and time-consuming data collection. This engine leverages user data, AI capabilities and client-specific requirements to create highly realistic synthetic personas and their responses to survey questions. Here's a detailed explanation of the process we undertook for Greencore and M&S.

Step 1: Data collection and initial profile building

The process began with collecting data from our platform users. This data encompassed various attributes such as demographics, behavioural patterns preferences, and interactions. The objective was to build a foundational profile for synthetic personas.

User data aggregation:

- Collection of data points from user interactions on the Bulbshare platform;
- Inclusion of demographic details (age, gender, location), behavioural data (purchase history, browsing patterns) and preference indicators (likes, dislikes).

Profile construction:

- We aggregated these data points to create a detailed profile for each synthetic persona;
- We ensured profiles were diverse and covered a broad spectrum of the target audience.

Step 2: Leveraging LLM capabilities for profile enhancement

With the initial profiles in place, we enhanced them using large language models (LLMs). The LLMs brought a layer of depth and nuance to the profiles, making them more reflective of real individuals.

Contextual enrichment:

- We used LLMs to analyse the collected data and generate additional context for each persona;
- We created detailed descriptors that include personality traits, likely opinions and potential responses to various scenarios.

Synthetic persona description:

- We formulated comprehensive persona descriptions that served as the basis for generating survey responses;
- We ensured these descriptions encapsulated the intricacies and nuances of real user profiles.

Step 3: Incorporating client requirements

Client requirements were critical in fine-tuning the synthetic personas to align with specific research objectives. This step ensured that the generated responses were tailored to the client's needs.

Requirement gathering:

- Engagement with the client enabled us to understand their specific research goals, target demographics, and any particular attributes of interest;
- We collected detailed client input on the desired characteristics of the synthetic personas.

Persona adjustment:

- We adjusted the synthetic personas based on client requirements;
- We fine-tuned attributes such as age, income level and interests to match the research objectives.

Step 4: Generating synthetic responses

With enriched and client-tailored synthetic personas, the next step was to generate survey responses.

Contextual input for AI models:

- We fed the detailed persona descriptions into the AI models;
- We used the contextual information to guide the AI in generating responses that were consistent with the persona profiles.

Response generation:

- Generation of survey responses.

Step 5: Validation and fine-tuning

The generated synthetic responses were compared with real-world responses to ensure accuracy and realism.

Comparison and analysis:

- We compared synthetic responses with actual user responses from similar surveys;
- We identified discrepancies and areas for improvement.

Fine-tuning:

- We adjusted the AI models and synthetic personas based on the comparison analysis.

Ethical considerations

Ethical considerations are paramount. We have the following compliance strategies for synthetic data usage across all clients.

Transparency in data generation:

- We clearly communicate to clients and stakeholders that synthetic data is being used;
- We provide detailed documentation on the methods used to generate synthetic data, ensuring that the process is understandable and replicable.

Ensuring privacy and confidentiality:

- Since synthetic data is used to avoid privacy issues; we guarantee that no real-world individual can be identified from the data;
- We regularly audit synthetic datasets to ensure that they are devoid of any real personal identifiers.

Maintaining professional responsibility:

- We use synthetic data responsibly to avoid misleading clients or stakeholders;
- We ensure that synthetic data is used to complement, not replace, real-world data where accuracy and realism are paramount.

Ensuring data accuracy and validity:

- We validate synthetic data against real-world data to ensure that it accurately reflects the patterns and distributions of actual data;
- We continuously refine synthetic data generation processes to improve the realism and utility of the data.

Obtaining informed consent:

- While synthetic data typically does not require direct consent from individuals, we inform stakeholders about its use and implications;
- In cases where synthetic data is derived from real-world data, we ensure that the original data collection was conducted with proper consent and ethical considerations.

Ethical use and avoidance of harm:

- We use synthetic data in a way that avoids harm to any individual or group, especially in sensitive research areas;
- We avoid using synthetic data to mask or justify unethical practices or biases.

We also adhere to and follow the guidelines set down by both ESOMAR and MRS in terms of code of conduct (see Appendices).

Our hypotheses

When conducting a market research study using both real-world (human) and synthetic (AI-generated) samples, the similarities and differences in responses between these two groups can be quite pronounced. For the M&S study, our hypotheses of how each type of sample might perform across different types of questions and techniques are outlined below.

Similarities

- **Fact-based questions:** For questions that require factual knowledge or straightforward information retrieval, both real-world and synthetic samples can provide accurate responses. For instance, queries about demographic information, product specifications or historical data will yield similar results.
- **Ranking/list of answers:** When asked to rank or list items based on specific criteria, synthetic samples can mimic logic used by real humans if they are trained on relevant and diverse datasets. AI can replicate patterns observed in human preferences quite effectively in these cases.

Differences

- **Prompted/open-ended questions:**
 - a) **Real-world sample:** Humans are more likely to provide nuanced and contextually rich responses, considering personal experiences, emotions and social contexts;
 - b) **Synthetic sample:** AI can generate plausible responses based on patterns in the training data, but it may lack depth in understanding personal context or unique human experiences. Responses might be overly generic or miss subtle human insights.
- **Projective techniques:**
 - a) **Real-world sample:** Humans excel at projective techniques (e.g., completing sentences, storytelling or interpreting ambiguous stimuli) because these methods tap into unconscious thoughts and emotions. Responses are often rich in subjective detail;
 - b) **Synthetic sample:** AI can attempt to simulate human-like projections, but it may fall short in terms of creativity, emotional depth and the ability to generate truly novel insights. Its responses are based on learning patterns rather than genuine unconscious processes.
- **Emotionally related questions:**
 - a) **Real life sample:** Humans provide responses based on real emotional experiences, which can be deeply personal and varied. This leads to rich, authentic data that reflects genuine human sentiment.
 - b) **Synthetic sample:** While AI can analyse and mimic emotional language to some extent, it lacks the capacity for genuine emotion. Its responses might be technically accurate, but may not capture the true depth and variability of human emotional experiences.

Our approach

Our study involved two parallel datasets: one comprising real-world sample from the Bulbshare and Greencore “Talking Taste” Community and including M&S sushi buyers, and the other consisting of synthetically generated data designed to replicate real-world sample attributes. The real-world data was collected by reaching out to the Talking Taste Community, and the synthetic data was generated using machine learning techniques trained on demographic and shopper patterns observed in the initial dataset. The core of our research lay in a comparative analysis of the real-world and synthetic samples. To gather our real-world data, we launched a survey to the talking Taste Community, which members completed on their smartphones via the Bulbshare app. The survey comprised two briefs:

1. General attitudes to sushi;
2. Views on sushi in M&S.

In Brief 1, we looked at general associations with sushi, the motivations and barriers to purchase, pre-packed versus made-at-counter sushi, the stereotypical sushi buyer and views on specific retailers including M&S and the ideal sushi box. In Brief 2, we focused on the sushi offer in M&S—i.e., what is done well, what could be improved, image associations with sushi in M&S, including the personification of the stereotypical M&S sushi buyer and the ideal M&S offer. These briefs were replicated and launched to the synthetic sample as a single mirror survey. Across both briefs we used a mix of closed and open-ended questions, including emotional associations and projective techniques. Please note that we currently only have the capability to ask open or closed questions to the synthetic sample, whereas we would ordinarily be able to ask a number of other question types to the real-world sample via the app, for deeper insights and enhanced user-experience (e.g., image highlighter tool, ranking, photo upload, video upload and swipe). This shaped and limited the survey as we had to ask the same questions of both the real-world and synthetic sample.

Our findings reveal both strengths and limitations in the use of synthetic data. Initial results indicate that the synthetic data can closely approximate the behavioural profiles of real-world sample, and that it performs well on fact-based answers and responding to lists, providing scale and granularity. However, there are minor anomalies where it puts emphasis on an attribute that is not so significant in the real-world data. In terms of open-ended questions, it has a tendency to be generic in some answers, or offering soundbites and cliches, and replaying back the typology in a literal way that we do not see in the real-world sample. It can, however, be eloquent and performs better on emotional and projective questions than anticipated.

The results: Synthetic data versus real life sample data

Fact-based and list questions

The synthetic sample performed well on the fact-based and list questions in terms of how it answered the questions, replicating to a degree the answers that we procured from the real-world data. Thus, for example, if 44% of participants in the real-world sample (base: 227) stated that they were interested in sushi generally, this was broadly reflected in the synthetic sample with 42% of the synthetic sample (base: 150), saying that they were generally interested in sushi (see Figure 1). Similarly, when looking at a list of barriers to eating sushi, the synthetic sample reflected the real-world sample when it came to the top barriers, with both citing expense, not liking raw fish and it not being very filling as the three key reasons not to buy sushi (see Figure 2).

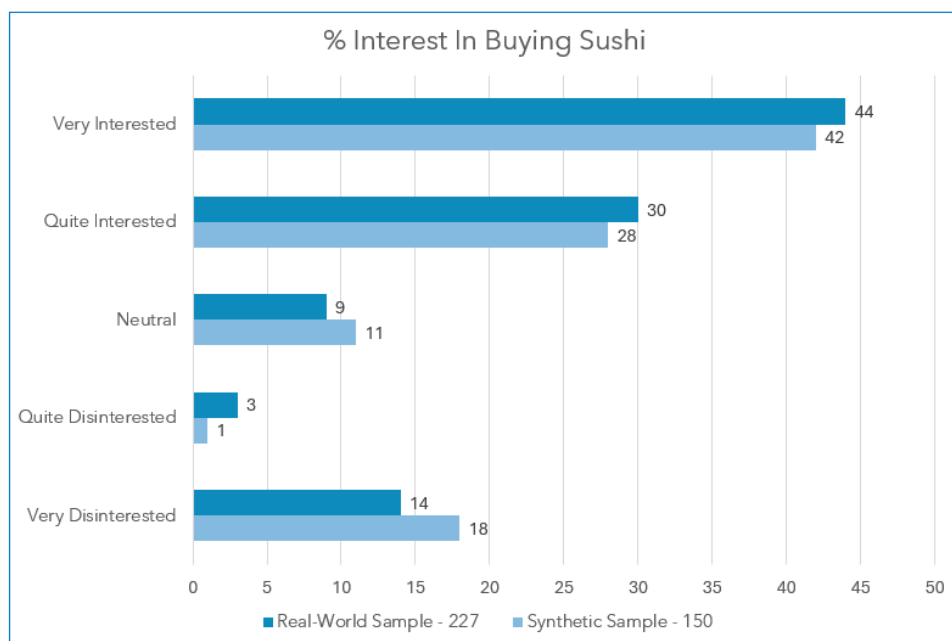


Figure 1. Interest in buying sushi.

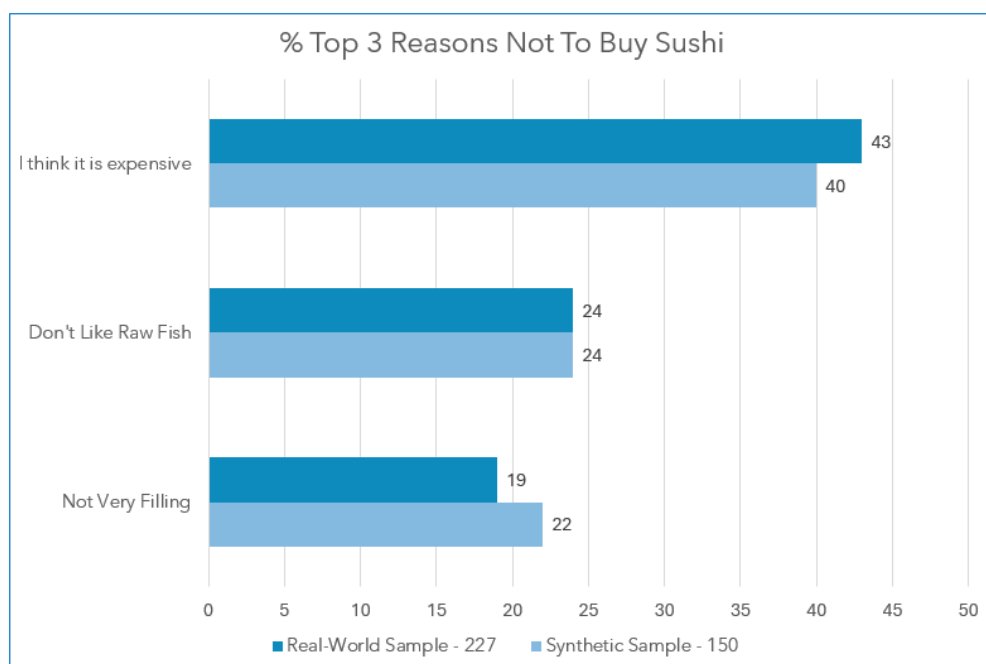


Figure 2. Top three reasons not to buy sushi.

However, we found in the first launch of the survey to the synthetic sample that it appeared to have a mind of its own. When presented with a list to respond to, or a single answer closed-ended question, rather than just indicating a response from the list, it added its own additional statements and viewpoint.

Example 1

When we asked the synthetic sample levels of interest in buying sushi with a single closed response on a five-point scale (very interested, quite interested, neutral, quite disinterested and very disinterested), we found

that the scores replicated the real-world sample as indicated above, but it added detail over and above the single response.

- “Very interested. Sushi presents a unique culinary adventure, a blend of flavours and textures that never ceases to amaze!” Curious foodie (synthetic sample).
- “Very interested. Sushi's a trendsetter, visually appealing, perfect for the 'gram. Plus, it's delicious”. Social statement-maker (synthetic sample).
- “Very disinterested. Prefer home-cooked meals, stick to my meal plans”. Homely planner.
- “Very disinterested. Value for money is my priority, sushi doesn't quite fit the bill”. Savvy budgeter (synthetic sample).

Example 2

We saw a similar expansion in response to the question when we asked the synthetic sample to indicate which supermarkets it buys sushi from currently and provided a list. Responses gathered included:

- “Frequent shopper at ALDI and Tesco. Good deals there. Haven't bought sushi elsewhere”. savvy budgeter (synthetic sample).
- “Wholefoods Store, Local Independent. These places often prioritize sustainability”. Planetarian (synthetic sample).

What the synthetic sample appeared to be doing, was adding in the attributes of the Greencore consumer typology (e.g., emphasising sticking to a list for the homely planners or VFM for the savvy budgeters) rather than just ticking the stores they buy from. We tweaked and resolved this for the second and third iterations, where the synthetic sample did what was asked of it and simply indicated the level of interest and the stores it purchased sushi from. Thus, looking at our hypotheses for how a synthetic sample would perform on fact-based and a list of answers to questions, it would seem that the hypotheses are validated in that it provided accurate responses and was able to mimic logic, even if in the first iteration it supplied more information than requested by the question. Perhaps it was a case of the synthetic sample wanting to be a people-pleaser by going that bit further! Or rather, by default it was drawing on the segments and typologies included in the original sample build.

Prompted/open-ended questions

When it came the open-ended questions and those that include projective techniques or emotions, we found some interesting observations. For some of the questions, the synthetic sample seemed to emphasise an element that did not have as high incidence in the real-world sample. For example, when we asked the synthetic sample how much price affects the decision to buy sushi, it replicated the real-world sample in terms of stating that quality and freshness was important, and sometimes more so than price; i.e., you get what you pay for. However, it was much more likely than the real-world sample to state that authenticity was important, with many more mentions of this. Also, when we asked about whether sushi purchase is more of a planned or impulse purchase, while the real-world sample was more likely to state that it was either more of an impulse buy or a mixture of impulse and planned depending on mood, craving and how fresh/appetising the product looks, the synthetic sample was much more likely to say that it was always a planned purchase. Further, the real-world sample was more likely to bring in nuance (role of offers, visual

appeal and role of prepacked) than the synthetic sample, which again drew heavily on the literal interpretation of the segments and typologies.

Real-world sample

- “It tends to be a bit of both so sometimes I buy it planned as a lunch meal as I am craving it and other times if I see it whilst shopping, I will buy it from the supermarket”. Overwhelmed solution seeker.
- “It is largely impulse as I go by how fresh it looks and if it is on offer”. Savvy budgeter.
- “It's mostly on impulse because it may be on a deal in the moment”. Homely planner.
- “Most of the time it is impulse buy as the sushi on display really catches your attention”. Planetarian.
- “A mixture but mainly impulse and appearance of sushi works best, the more fish the better. Fresh sushi is planned. Pre-packed tends to be impulsive”. Curious foodie.

Synthetic sample

- “Definitely planned. Surprise sushi? No, thank you. I need to know what I'm eating and when. I make a list, stick to it. It's all about control, you see. I like my meals predictable and familiar. Plus, I'm not much for takeouts. Home-cooked meals are my thing!” Happily healthy.
- “Planned, definitely. Sushi's a treat, not an everyday thing. Gotta stick to the budget, you know? I watch for deals, specials. Even better when I find a coupon. I get more bang for my buck that way”. Savvy budgeter.
- “Planned, always. I need to ensure it's sustainably sourced. Can't support overfishing or environmentally damaging practices. It's about the planet's health, not just the taste”. Planetarian.
- “Planned, absolutely. Sushi is a treat, not an everyday meal. Got to ensure it fits into my meal plan for the week. Plus, I like to know where it's coming from, freshness is key. Can't risk a last-minute purchase messing up my schedule or my stomach!” Homely planner.

Projective techniques and emotionally-related questions

Looking at projective and more emotionally-related questions, we were pleasantly surprised to see that the synthetic sample performed rather well. It was able to “imagine” and provide detailed and rich responses.

Example 1

In response to the question “How do you feel about sushi?” responses from the synthetic sample were on a par with the real-world sample, although again, heavily shaped by the segment profile:

- “Sushi? Not really my cup of tea. I prefer sticking to my home-cooked meals. The comfort, the familiarity, it's irreplaceable. Plus, I like to know what's going into my food, and with sushi, it's a bit of a mystery”. Homely planner (synthetic sample).
- “Sushi? Love it! It's a whole art form on a plate. The combo of fresh fish, perfectly seasoned rice, and that hint of wasabi? Divine! Always on the hunt for new sushi spots”. Curious foodie (synthetic sample).

- “Oh, sushi! Big fan. Love the diversity in flavour and ingredients. Makes for a fun dining experience, you know? Trying out different types and sharing on social media, that's my thing. But, finding authentic sushi, that's a challenge sometimes”. Social statement-maker (synthetic sample).
- “Love it! Always up for trying different sushi rolls. The mix of flavours, it's a whole new experience each time. Can't resist a good California roll!” Curious foodie (synthetic sample).

Example 2

In response to a personification, answering the question, “If M&S Own Label Sushi came to life and was a real person, what type of person would they be and what would they be like?” we found some rich descriptions and ones that were more detailed than many of the real-world answers. However, there was also a lot of repetition of specific characteristics (e.g., meticulous, attention to detail, minimalist, perfectionist) so that some of the answers looked like clones, which was not evident in the real-world sample where there was more diversity and nuance.

“M&S Own Label Sushi, right? I'd say it'd be a person of elegance and simplicity. No fuss, just quality. They'd be the type who values tradition, respects the art of precision. Someone who's meticulous, much like a sushi chef. They'd be quiet, but their presence would command respect. They'd appreciate the small details in life, just like how every grain of rice matters in sushi. A person of few words, but when they speak, it's meaningful. They'd be a perfectionist, always striving for the best. Imagine them in a well-pressed suit, always looking sharp. That's M&S Own Label Sushi” (synthetic sample).

“M&S Own Label Sushi, if it were a person, would be meticulous, detail-oriented, and conscious about quality. They'd be someone who values tradition but isn't afraid to put a modern twist on things. Health-conscious, they'd be a stickler for fresh, natural ingredients. They'd embody the same discipline and precision that goes into sushi making, reflecting the balance and harmony of flavours in their own life. They'd be the type who's always striving for perfection, but also appreciates the simple, clean aesthetic” (synthetic sample).

“M&S Own Label Sushi, huh? Interesting. I'd say it'd be a person who's meticulous and detail-oriented. They'd have a minimalist style, sleek and modern. They'd value tradition but also embrace innovation. They'd be a bit reserved, but once you get to know them, there's a depth of character and complexity. They'd probably be into arts, like calligraphy or origami, and enjoy quiet pastimes like reading. They'd be health-conscious and environmentally aware. A bit of a perfectionist, you know? Always striving for balance and harmony in life” (synthetic sample).

“Hmm, M&S Own Label Sushi...Let's see. They'd be a minimalist, definitely. Clean, simple, no frills—just like sushi. Probably a bit of a perfectionist too, caring about details. Not too flashy or extravagant. Quality over quantity, that's their mantra. They'd be disciplined and methodical, just like sushi-making. And they'd probably be a bit of a health nut too, given sushi's low-cal, high-protein profile. But hey, they'd also be a bit adventurous—sushi's all about trying different things, right? So, a healthy, disciplined, adventurous minimalist. That's M&S Own Label Sushi personified” (synthetic sample).

“It would be a bright bubbly sophisticated woman who cared about people and their health and was generous” (real-world sample).

“People who look for healthy options but don't want to compromise on taste. They enjoy an active lifestyle and also enjoy the finer things in life” (real-world sample).

“A sophisticated middle-aged person with a great job, popular with lots of friends and a great social life” (real-world sample).

“A fun, vibrant and health-conscious individual who is slim and middle class” (real-world sample).

“It would be someone well-travelled and knowledgeable about other cultures. Also, someone who likes quality and doesn't mind paying more” (real-world sample).

Therefore, a bit like Schrödinger's cat, our hypotheses on open-ended/projective and emotionally-based responses are both validated and invalidated. We were able to gather rich and detailed responses from the synthetic sample-making for the basis of good insights. However, we also see that the synthetic sample relied heavily on the algorithms, so that we got less diversity of response, more conformity and repetition and less nuance. The synthetic responses spoke in a way that seems more contrived than the natural, real-world responses. The reason for this may be based in the following considerations:

- Training data limitations: Synthetic responses are generated based on patterns found in the training data. If the training data lacks diversity or contains biases, the generated responses will reflect these limitations. The training data may not cover the full range of possible human experiences, leading to less varied responses.
- Algorithmic constraints: The algorithms used to generate synthetic data often rely on statistical methods and predefined structures. These methods can struggle with the nuance and creativity required for open-ended questions, leading to repetitive or formulaic responses.
- Context understanding: Generating diverse and contextually appropriate responses requires a deep understanding of the subtleties of language and context. Current AI models can struggle with fully grasping the context of open-ended questions, leading to answers that may not fully address the query or explore its nuances.
- Safety and appropriateness filters: AI systems often have filters to prevent the generation of inappropriate or harmful content. These filters can overly constrain the responses, making them sound less natural and more generic.
- Overfitting to common patterns: AI models tend to generate responses that align with common patterns and phrases found in the training data. This can lead to responses that lack originality and appear contrived, as the model favours safe, well-trodden linguistic paths over novel or less common expressions.
- Lack of personal experience: Human responses to open-ended questions are often rich with personal experiences, emotions and individual perspectives. AI models do not have personal experiences or emotions, which can make their responses seem less authentic and more mechanical.

Bulbshare seeks to increase the diversity and naturalness of synthetic responses by addressing these challenges where possible.

Additional observations: US bias and tone of voice

There are two additional points to consider when we look at synthetic data versus real-world data in the M&S sushi study.

Bias

When we conducted our Greencore/M&S Sushi study, the responses generated were very US in tone rather than UK-oriented, which is where the Greencore Talking Taste Community is based. In response to the question: *Why this level of interest in sushi?*

“Interest? It's all about the dollars and cents, friend. You see, every penny saved is a penny earned. I'm always on the hunt for the best bang for my buck, and I've become quite skilled at stretching my food budget”. Synthetic sample.

When we asked *Who does the best sushi and why?*, we received repeated mentions from the synthetic sample of a New York sushi restaurant that did not figure once in the real-world sample. Additionally, there were relatively few references to UK supermarkets:

“Sushi Yasuda in New York. Precision, quality ingredients and experienced chefs. Rice texture is perfect. Fish, always fresh. Traditional sushi at its best. No fancy rolls, just expertly crafted nigiri. Worth every penny” (synthetic sample).

“Sushi Yasuda, New York. Meticulous, like me. Chefs there, true craftsmen. Each piece, a work of art. Fresh fish, perfect rice, balance of flavours. Tradition, simplicity, perfection. That's why” (synthetic sample).

“Sushi Yasuda in New York, hands down. They use the freshest ingredients. Their sushi chefs are artisans, really. Each piece is crafted with precision, balance of flavours is impeccable. Plus, their rice, it's slightly warm, contrasts beautifully with the coolness of the fish. It's all about the details, you know? And they nail it, every single time” (synthetic sample).

“Best sushi? Difficult to pinpoint. Sushi Yasuda in New York, top-notch. Quality of fish, superb. Rice, just right. Chefs, highly skilled. Aesthetics, simple yet elegant. Pricey? Yes. Worth it? Absolutely” (synthetic sample).

Tone of voice

We found that the style and tone of voice of the synthetic sample was really quite exuberant, florid and prone to hyperbole and cliché in some instances, in a way that the real-world sample was not, being UK-based and perhaps a bit more tonally measured. In response to the question *Why this level of interest in sushi?*, the synthetic sample offered:

“Spontaneous, mostly! Sushi's like a treasure hunt, never know what unique flavour combo you'll stumble upon. It's not just about the food, it's the thrill of discovery! Keeps the palate on its toes, you know. But, if there's a new sushi joint in town, I'll plan a visit. Got to keep my sushi game strong!”

“Health is wealth! Keeping the body in good shape is a priority. It's all about understanding the link between diet and wellness. Misleading food labels? Not a fan. The hunt for genuine, unprocessed foods can be a struggle, but worth it. It's not just about living, it's about living well”.

Meanwhile, the real-world sample stated:

- “This level of interest, because it's delicious and healthy”;
- “I really don't like the thought of eating raw fish”.

Moreover, in response to the question: *Whatever you feel about sushi, if you could create your ideal sushi box, what would be in it and why?*, the synthetic sample stated:

“Ah, the sushi box! A canvas for culinary creativity, no doubt. I'd start with a base of perfectly seasoned sushi rice, of course. Then, let's bring in the sea's bounty—a mix of sashimi-grade tuna, salmon and yellowtail. Can't forget the ebi, cooked shrimp, for a touch of sweetness. Now, let's get adventurous! How about some uni, sea urchin? It's a delicacy, not for everyone, but I find its rich, briny flavour fascinating. A few pieces of unagi, grilled eel glazed with a sweet soy-based sauce, would add a nice contrast. For the veggie lovers, I'd include some avocado, cucumber and pickled radish rolls. They offer a refreshing crunch that balances out the rich seafood. Lastly, a sprinkle of sesame seeds for a nutty finish, and a side of pickled ginger and wasabi to cleanse the palate between bites. And there you have it, my ideal sushi box! A symphony of flavours and textures, a journey around the world in every bite”.

On the other hand, this is what we got from the real-world sample:

- “A selection of three to four pieces featuring different fish”;
- “A mix of classic and innovative rolls, such as California roll, spicy tuna roll and vegetarian options like avocado and cucumber rolls”;
- “Two to three slices of pure, high-quality fish”;
- “Sweet tofu pockets filled with seasoned rice”.

Implications and conclusion

The adoption of synthetic data in market research is still evolving, and our study indicates that there are strengths to it, where it can mimic real-world data accurately so is ideal for scale when it is a challenge to get a robust real-world sample, or when highly confidential material (e.g., new products) is to be shared. However, there are limitations, where it can be generic or relatively constrained as it reflects algorithms rather than the true human psyche. As we hone and refine our use of synthetic samples and accompanying data, our view is that future developments might include:

- **Advanced algorithms:** Continued advancements in machine learning and artificial intelligence will improve the quality and realism of synthetic data, making it an even more powerful tool for market researchers.
- **Personalised marketing strategies:** With more granular data, businesses can develop highly personalised marketing strategies tailored to the unique preferences and behaviours of niche consumer segments—in this case M&S sushi buyers. This targeted approach can enhance customer engagement and loyalty, driving better business outcomes.
- **Risk mitigation in market entry:** For companies considering entry into niche markets, synthetic data can serve as a valuable tool for scenario analysis and risk assessment. By simulating market conditions and consumer responses, businesses can make informed decisions and develop strategies that mitigate potential risks.

Our final thought is that the development of sophisticated models that seamlessly integrate synthetic and real data will enhance the robustness and validity of market research findings. We are excited to be part of the synthetic data journey and look forward to integrating it more as we learn and develop with it.

About the authors

Julia Brannigan, Senior Manager, Bulbshare, London, UK.

Kerry Jones, Insight Controller, Greencore, Worksop, UK.

Appendices

By adhering to the ESOMAR and MRS codes of conduct, we can leverage synthetic data effectively and ethically. Ensuring transparency, maintaining privacy, upholding professional responsibility and guaranteeing data accuracy are fundamental to this process. Compliance not only protects from reputational risks, but builds trust with our clients such as Greencore and M&S, and fosters a sustainable and ethical research environment.

Appendix 1: ESOMAR

Transparency:

Researchers must clearly disclose when synthetic data is used.

The creation and use of synthetic data should be explained, including its purpose and how it was generated.

Data quality:

Synthetic data should meet the same quality standards as actual data.

It must be accurate, relevant and fit for the intended research purpose.

Privacy and confidentiality:

Synthetic data should be used to protect the privacy and confidentiality of individuals.

Researchers must ensure that synthetic data cannot be reverse engineered to identify individuals.

Informed consent:

When synthetic data is created from real data, the original data subjects must have given consent for their data to be used in this way.

Researchers must be transparent with data subjects about how their data will be used.

Ethical use:

Synthetic data should not be used to mislead or deceive.

It must be used ethically and in a manner that respects the rights of all stakeholders.

Appendix 2: MRS

Integrity:

Researchers must not knowingly use synthetic data in a way that misrepresents findings.

They should strive for honesty and accuracy in reporting research results derived from synthetic data.

Transparency:

The use of synthetic data must be disclosed in research reports.

Information about how synthetic data was generated and its limitations should be provided.

Data protection:

The generation and use of synthetic data should adhere to data protection laws and regulations.

Researchers must ensure that synthetic data cannot be traced back to real individuals, maintaining anonymity.

Consent:

Similar to ESOMAR, if synthetic data is based on real data, informed consent from original data subjects is necessary.

Participants should be aware of how their data might contribute to the creation of synthetic data.

Professional responsibility:

Researchers should use synthetic data responsibly, ensuring it is appropriate for the research objectives.

They should also consider the implications of using synthetic data on the validity and reliability of their research.