IPSOS VIEWS

THE POWER OF PRODUCT TESTING WITH SYNTHETIC DATA

Humanizing AI series, part two

Colin Ho, Ph.D
Dr. Nikolai Reynolds

At Ipsos, we champion the unique blend of Human Intelligence (HI) and Artificial Intelligence (AI) to propel innovation and deliver impactful, human-centric insights for our clients.

Our Human Intelligence stems from our expertise in prompt engineering, data science, and our unique, high quality data sets – which embeds creativity, curiosity, ethics, and rigor into our AI solutions, powered by our Ipsos Facto Gen AI platform. Our clients benefit from insights that are safer, faster and grounded in the human context.

#IpsosHiAi

> At Ipsos, we believe synthetic data opens brand new possibilities for market research, particularly in the field of product testing.

Synthetic data is about to change the world. From fast-tracking drug development in healthcare, simulating fraudulent transactions in financial services, and fueling autonomous vehicle tests in the automotive sector, it is already demonstrating its value across various business contexts.

At Ipsos, we believe synthetic data opens brand new possibilities for market research, particularly in the field of product testing. However, many businesses remain uncertain about the quality of synthetic data, or how to evaluate it. This paper aims to bridge these gaps.

As the world's largest and leading product testing adviser, Ipsos has been at the forefront of leveraging cutting-edge technologies to accelerate innovation and growth for businesses around the world. The following pages present Ipsos' insights of testing products with synthetic data, providing readers with:

- Recommendations for generating and evaluating high-quality synthetic data sets
- Specific applications in product testing for consumer goods and services

Today, we find different types of synthetic data in the market research industry, each with their strengths and weaknesses. In this paper, we focus on data augmentation, i.e., enhancing datasets with synthetic data.

Types of approaches to using synthetic data

Data augmentation

Enhancing datasets with synthetic data to create a more comprehensive sample, while maintaining statistical integrity

Data imputation and fusion

Filling in missing data points using existing information

Gen AI agents and persona bots

Tailored digital assistants that mimic consumer segments, offering insights from synthesized responses

Full synthetic data

Utilizing entirely artificial samples made up of synthetic respondents

We start with a general overview of what is needed to generate high-quality synthetic data and how to evaluate its quality. As the effects and nuances of synthetic data can only truly be understood when considered in specific application areas, we investigate how synthetic data can be applied to product testing, specifically. This paper covers only the generation of synthetic *numerical* data; the format used most by quantitative researchers. It does not cover the application of synthetic images, video, data imputation, or synthetic personas, all of which also fall under the broad umbrella of synthetic data[1].

As explained in the Ipsos Views paper, Synthetic Data: From Hype to Reality[1], synthetic data is artificial data that is generated from a model that is trained to mimic the statistical properties and patterns of real-world data.

> If an AI has not been trained on real-world data that is *relevant* to your business, it will not be able to generate synthetic data that shares the same properties as real-world data. It's as simple as that!

Generating and evaluating synthetic data

We use data to make better business decisions in the real world, and synthetic data can be employed in many ways to support decision-making. As such, while synthetic data does not correspond to real events or people, it still needs to mimic the statistical properties and patterns of real-world data. This raises two fundamental questions:

01 What is required to generate synthetic data that closely mimics real data?

02 How can synthetic data be evaluated for its resemblance to real-world data?

Before an AI can generate synthetic data that mirrors real-world data, **an AI needs to be trained on real-world data.** As discussed in Ipsos' first paper in the Humanizing AI series, AIs are simply algorithms; they have no intelligence of their own, until they are trained. It is through learning from training data that AIs gain the intelligence we associate with them. This is the most critical point to remember from this paper: if an AI has not been trained on real-world data that is *relevant* to your business, it will not be able to generate synthetic data that shares the same properties as real-world data. It's as simple as that!

The evaluation process is straightforward as well. Synthetic numerical data should, at minimum, **mirror real-world data on common statistical measures**— such as means, data distributions, variances, and relationships between variables (e.g., correlations). A direct comparison between synthetic and human data on these common metrics will provide us with a sense of how well a synthetic data set approximates human data. The closer synthetic data is to human data, the less risk we assume when using it, but there is *always* some risk because synthetic data can never perfectly mimic real data in every aspect. We should use synthetic data, therefore, only when we are willing to accept some risk.

Generating synthetic data using LLMs

Approaches to generating synthetic data can be divided into two categories: LLMs (Large Language Models) and non-LLMs, differentiated by their text-based and numerical-based nature, respectively. Off-the-shelf, or public LLMs, which are pre-trained on extensive datasets such as websites, online books, and social media posts, can generate quality synthetic data in subject areas included in their training.

However, off-the-shelf LLMs have limitations in producing realistic synthetic data (see Figure 1)[2,3]. First, their training data is limited in coverage – many topics are too mundane or private to be found online. Second, LLMs are often biased toward Western, English-speaking countries, due to the predominance of such data in their training sets. Studies have shown, for example, that cultural values generated by LLMs align more closely with the Anglosphere and Protestant Europe than those of other countries[4]. Third, information can quickly become outdated.

Therefore, to generate high-quality synthetic data using LLMs, it is crucial to train them on updated, country-specific real-world data relevant to the subject of interest. This process requires access to recent, pertinent, and specialized data, statistical and data science expertise, and substantial investment of time and effort to ensure the synthetic data accurately reflects real-world statistical properties and patterns[5].

Figure 1: Comparisons between an all-human dataset and human-synthetic augmented dataset

Source: Ipsos

Generating synthetic data using non-LLM approaches

Long before LLMs stole the spotlight, data scientists have used Deep Learning (DL) algorithms[6] to generate synthetic numerical data[7]. DL algorithms, including the types used in LLMs, are potent tools for the generation of synthetic data, each possessing unique advantages.

LLMs are particularly effective at generating human-like text data. They can provide detailed and contextually rich textual data, making them highly valuable in applications such as content creation, language translation, and chatbots[8].

Non-LLM DL offers significant advantages for generating synthetic numerical data. DL algorithms are particularly effective at producing synthetic numerical data that closely mirrors the statistical properties of real-world datasets. While pre-trained LLMs like ChatGPT are designed for natural language tasks, DL models can be trained specifically for numerical data synthesis, enabling customization for domain-specific or market-relevant applications. At Ipsos, we have a strong history of leveraging DL techniques for data synthesis. In the following section, we detail the results of applying DL models to generate synthetic data for product testing, analyzed on a product-by-product basis.

Why product testing

Outside of market research, while many applications of synthetic data focus on the anonymity (e.g., anonymizing medical data to preserve confidentiality), in market research, the key benefit many businesses are looking for is cost and time savings from having to collect real-world data.

With online data collection getting faster and less expensive each year, one needs to carefully consider whether the savings in time and money are worth the decrease in accuracy that comes with synthetic data. For example, using Ipsos.Digital, Ipsos' agile testing platform, a researcher in the United States can conduct a survey of 300 respondents for around USD $2,000 and get results in around 24 hours. Hypothetically, if the researcher leverages synthetic data to answer the questions the 300 real humans would have provided, answers with synthetic data may take 12 hours to generate, cost $1,500, and offer lower accuracy compared to the real human data. In this situation, it may make more sense to collect real data rather than generate synthetic data as the savings in cost and time do not seem worth the decrease in accuracy. Why not spend the additional $500 and wait 12 hours longer to get the real thing!

Due to the inherent trade-off that comes with synthetic data, we wanted to test synthetic data in scenarios where the cost of conducting research is typically high. Product testing fits this bill perfectly due to the many costs involved:

- **Manufacturing:** Manufacturing or purchasing products and prototypes, or masking them for product testing
- **Shipping and Returns:** Costs to deliver products, and return or destruction of empty packaging
- **Sampling:** Recruitment costs, particularly when businesses need to be selective about their user base

Due to manufacturing, shipping, and sampling costs, any reduction in the number of participants in a product test can lead to significant savings. This does not mean that synthetic data cannot be used in other areas of market research, it simply means the risk may outweigh the benefits if cost savings are minimal.

We still need humans

The product experience is inherently human. AI alone cannot capture the five senses, emotions, expectations, or the impact of context that humans experience with products. Therefore, our goal in applying synthetic data to product testing was not to replace human input entirely, but to augment it. Our challenge was to establish the minimum number of human respondents needed to test products alongside synthetic data, to ensure viable results. To achieve this, Ipsos' Innovation teams ran two research streams:

Research stream 1

Determining the smallest number of humans needed to approximate product testing results from larger samples (e.g., 200-300 humans) *without synthetic data*

Research stream 2

Validating that small human sample, when augmented *with synthetic data*, yield the same results as all-human samples

In **research stream 1**, we leveraged some of the data from Ipsos' product testing database, considering 40,000 respondents and 185 selected consumer packaged goods (CPG) products tested worldwide[9], to assess at what minimum number of human samples we get a high enough correlation with product testing results from larger samples. We determined that when the best performing product differs from the worst-performing product by at least 8%, a sample of 50 human respondents is sufficient to replicate the performance rankings of the best and worst products (correlation coefficient r = 0.8).

The small number of participants needed to replicate results from larger samples is likely because variance in product testing data is mainly influenced by differences in product technology (e.g., the amount of sugar used). **Consequently, the variance observed in product testing is typically smaller than those in other areas of research, such as evaluating consumer attitudes or brand perceptions, for example.**

> The product experience is inherently human. AI alone cannot capture the five senses, emotions, expectations, or the impact of context that humans experience with products.

In an ideal scenario, businesses would act on these findings and conduct product testing with 50 participants, particularly during early-stage testing, when the risk is lower.

However, most businesses do not proceed in this manner for two reasons:

01 A sample size of 50 does not allow businesses to gain insights into subgroups. Sometimes, businesses need to analyze specific segments within the consumer population.

02 A sample size of 50 results in low statistical power for detecting differences. This is problematic, because businesses usually rely on action standards based on statistical testing. In practical terms, using a sample of 50 may prevent businesses from moving forward with the research findings. In statistical terms, small samples increase the risk of Type II errors—the chance of failing to detect a real difference when one exists.

Therefore, instead of recruiting 200 humans to test a product, for example, we could recruit 50 humans to test a product, generate 150 synthetic respondents from the 50 human data without replicating or resampling the 50 humans, and then combine the human and synthetic respondents to test with a mixed sample of 200. Consider the 50 humans as a "seed" sample to train the AI, enabling it to generate synthetic data that accurately mimics human responses to products.

This is where **research stream 2** comes in, to validate whether the small human samples, augmented with synthetic data, would still produce the same results as all-human samples. To ensure generalization, in our synthetic data pilots, we covered a diverse set of countries and categories. We also experimented with larger seed samples (e.g., 75, 100) but for brevity, will share only the findings from our trials with 50 humans.

Moving on, data from the 50 humans was used to train a DL algorithm to generate synthetic data. We did not use an off-the-shelf LLM because the public data that LLMs have been pre-trained on does not include people's multi-sensory experiences on products in the given category (e.g., prototypes, new formulations) and does not provide robust numerical data on respondent level. To add, we did not consider weighting the data as an alternative to DL, because weighting does not help in creating additional sample sizes for subgroups and is often not accepted in product testing.

> [Ipsos] validated results by comparing findings from an all-human dataset with those from a dataset augmented with synthetic data.

In general, synthetic data works

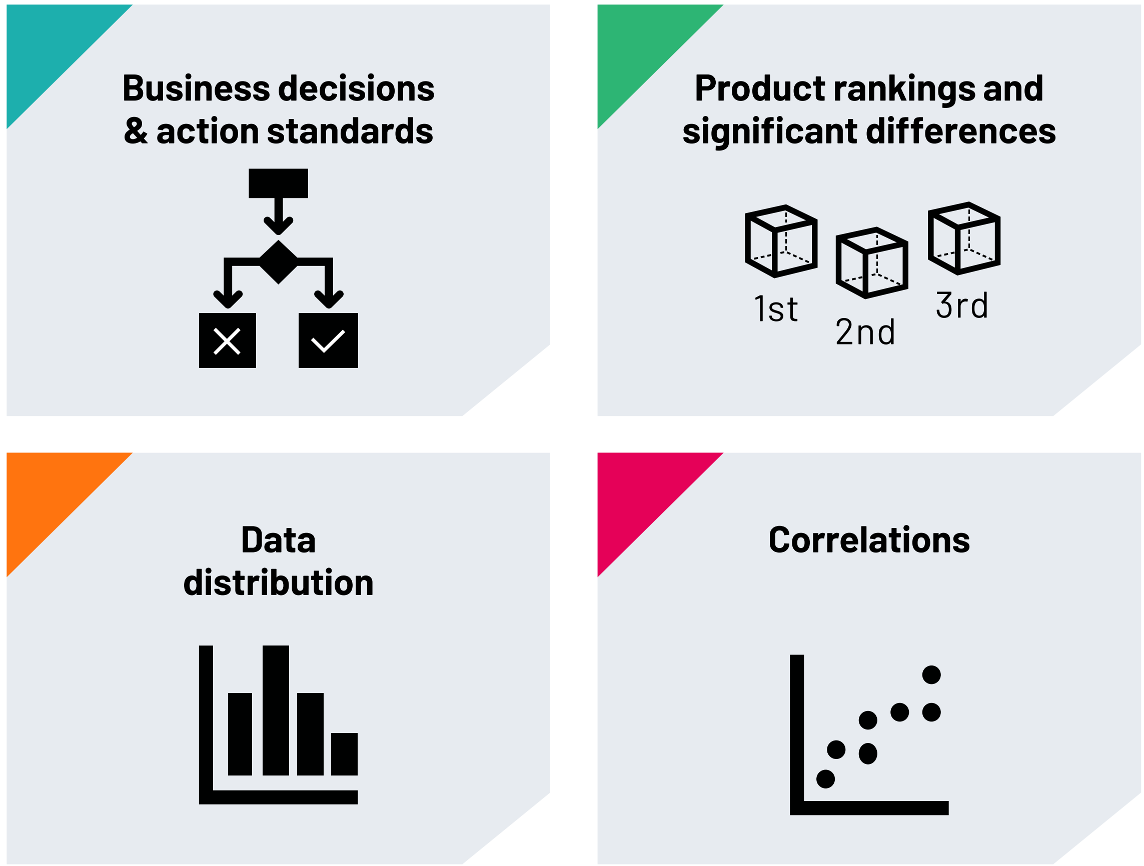
In our experiments, we validated results by comparing findings from an all-human dataset with those from a dataset augmented with synthetic data (see Figure 1)[10]. As a reminder, in our approach, we are not just replicating or copying and pasting existing respondent data.

In general, we found that the two datasets were remarkably similar, in in terms of:

- The relative performances of products (e.g., rankings, statistical significance)
- Data distribution (e.g., the distribution of people's responses across the answer options on individual questions)
- The relationships between variables in the data (i.e., correlations between overall liking and product attributes)

Most importantly, the two datasets showed differences in variances but led to the same business decision in every dataset we tested (see Figure 2).

Figure 2: Criteria used to determine accuracy of part-human, part-synthetic in replicating findings from the all-human dataset

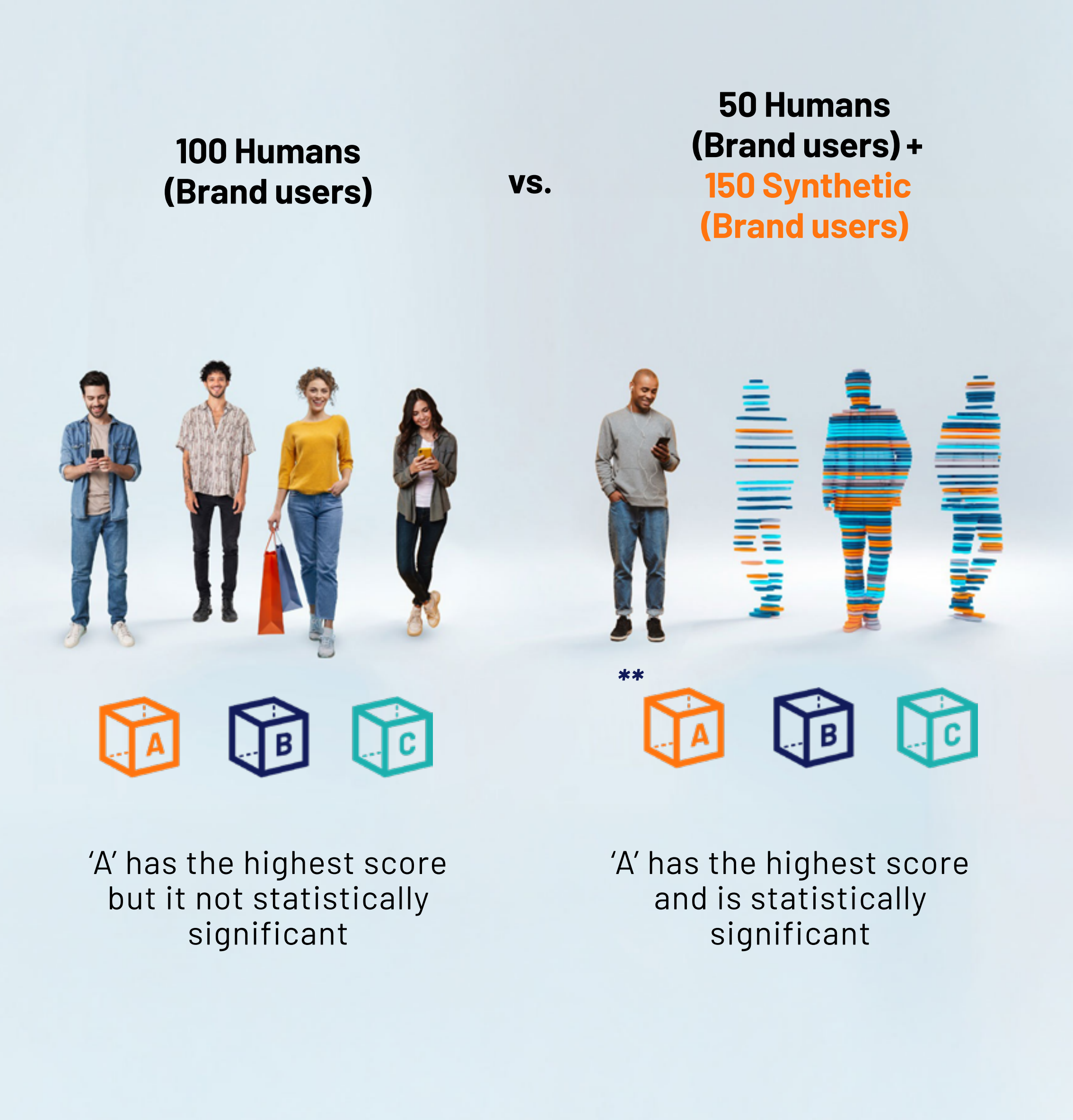

Business decisions & action standards
Product rankings and significant differences
1st
2nd
3rd
Data distribution
Correlations

Source: Ipsos

A key benefit of product testing with synthetic data is the ability to augment data for hard-to-reach populations. Once augmented, differences that were previously not statistically significant may become significant due to the boost in sample size. In one of our tests, for example, we generated synthetic responses to augment people who used a particular brand of product.

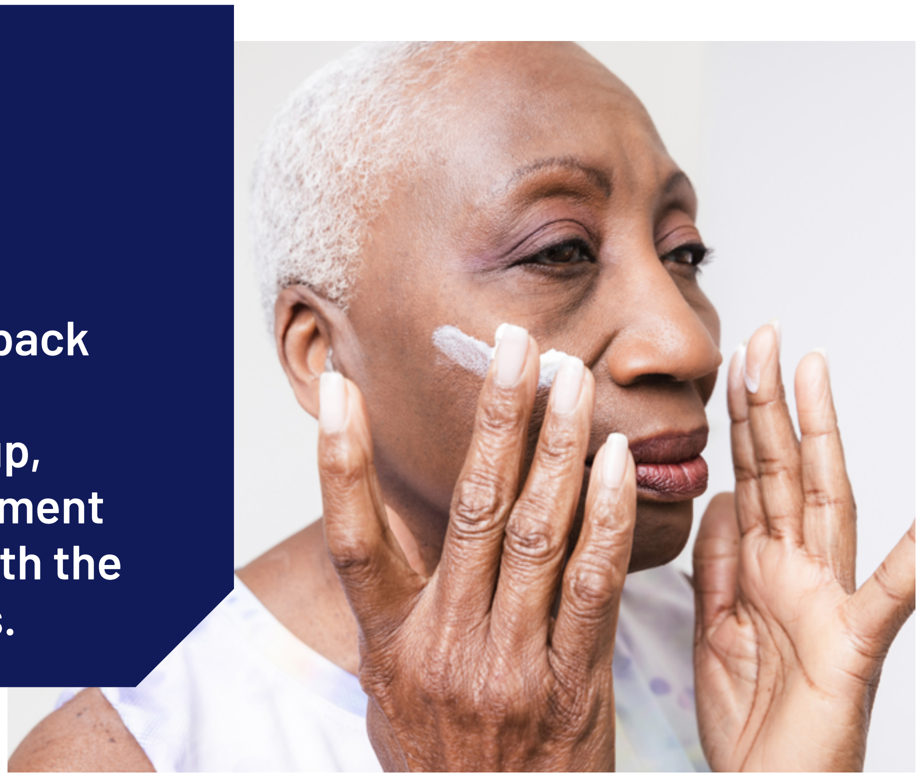
In our all-human sample, we had about 100 brand users per product. In the all-human dataset, there were some differences between three products tested, but the differences did not reach statistical significance. Once augmented with 100 synthetic brand users per product, the differences between products became statistically significant due to the increase in sample size (see Figure 3).

Figure 3: Human brand users augmented with synthetic brand users


100 Humans (Brand users)
vs.
50 Humans (Brand users) +
150 Synthetic (Brand users)
**
A
B
C
A
B
C
'A' has the highest score but it not statistically significant
'A' has the highest score and is statistically significant

Source: Ipsos

> When seeking feedback on products from a specific target group, we must set recruitment quotas that align with the business objectives.



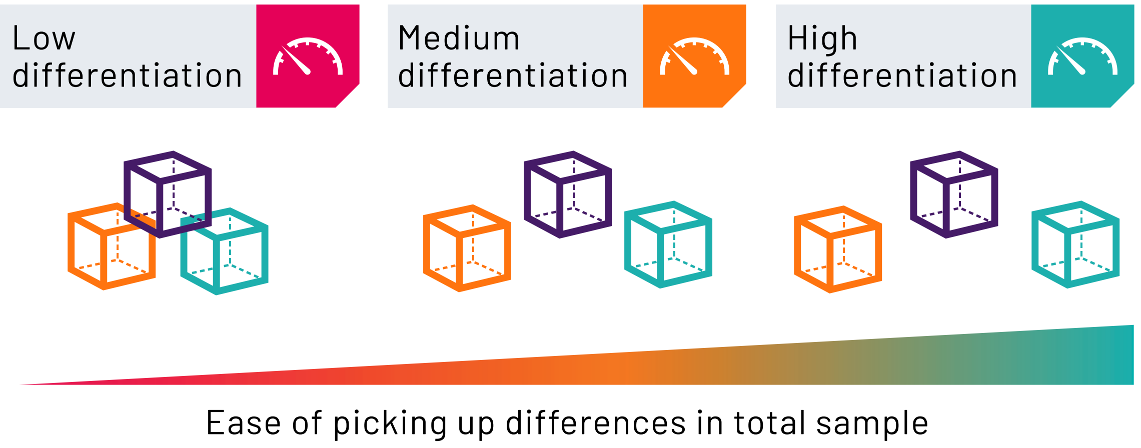
Some caution is warranted

We have presented a promising picture of synthetic data. As previously mentioned, synthetic data alone comes with an inherent trade-off, as it can never fully match the accuracy of real-world data. The ability of a part-human, part-synthetic dataset to replicate the findings of an all-human dataset depends on:

- **The representativeness of the human data used to train the AI.** In our initial pilots, when we did not control the 50 human participants to align with the structure of the original 200-person sample and generated synthetic data from the 50, we noticed differences in the results. Without such control, in 20-30% of these samples, the synthetic data generated did not capture the product performance reflected in the original 200 humans. The takeaway is that synthetic data can fail if the human seed sample is not representative. However, when we matched the structure of the 50 human participants with those of the original 200-person sample, the business decision would have been the same using a combined sample of 50 humans and 150 synthetics as it would have been using 200 humans. To summarize, when seeking feedback on products from a specific target group, we must set recruitment quotas that align with the business objectives. Just as we did in the past with larger sample sizes, this ensures we are sampling the right target group.

- **How differentiated the products are in the first place**. When there are significant differences between products in the all-human dataset, the AI was easily able to pick up those differences and replicate them in the synthetic data. When differences between products are not statistically significant, the AI, like the all-human sample, could not detect distinctions either (see Figure 4). In the numerous pilots we conducted across various countries and categories, we identified an additional advantage of using augmented samples in situations where there is low discrimination between products. When analyzing subgroups by dividing the original sample into smaller subsets, like current brand users and non-brand users, we discovered that using AI-generated synthetic data to enhance the smaller sub-group sizes improved the detection of differences in product performance, based on target group preferences. This was evident at the subgroup level, where the synthetic data helped increase the statistical power and uncover preferences that might not have been detected with the original sample alone.

Figure 4: Accuracy of synthetic data depends on actual product differences



Source: Ipsos

As a final note on validation, to avoid a reader leaving with the impression that they can start doing all market research with just 50 humans, we should make it clear that the findings presented here are specific to product testing. The conditions that made it possible to use only 50 humans in product testing may not be present in other areas of research. As the world's leading product testing adviser, we took some learnings from Ipsos' product testing database. Not to mention, product testing has unique data characteristics that can help ensure the quality of synthetic data (for instance, product characteristics that influence the human senses). Synthetic data should always be validated, on a case-by-case basis, for the specific application desired (e.g., segmentation).

> Data is often considered the lifeblood of businesses, enabling smarter decisions that help them grow and thrive.

The promise of synthetic data: from hype to reality

Data is often considered the lifeblood of businesses, enabling smarter decisions that help them grow and thrive. The promise of being able to generate synthetic data, at will, and at scale, is therefore extremely attractive. Public sentiments on synthetic data, however, are quite polarized. Companies offering synthetic data services might say "this is all you need, no humans required"! Researchers who are more cautious may adopt a wait-and-see approach and are hesitant to use synthetic data for the time being. Underlying these opinions is a tendency to categorize the world in a binary way: Good or bad? Synthetic or real-world data? We have taken a first step to provide clarity on this topic, to show that synthetic data is no different than other research tools. Synthetic data has its strengths and weaknesses and is more suited to certain situations.

In our pilots, we demonstrated that AI could generate synthetic data to mimic real-world data, but first, it requires quality human data for training purposes. Therefore, the answer is not synthetic or real-world data. We need both. The accuracy of synthetic data is not good or bad; rather, "it depends". If product differences are small among humans, we need to look into subgroups. If the human data used to train the AI is not representative for the target group or relevant to the business, then the accuracy of the synthetic data will be compromised. If we want to use synthetic data, we must accept that it may not work under some conditions. **As researchers, our responsibility is to ensure we use synthetic data only when appropriate: under conditions that will maximize success.** Augmenting synthetic data offers several advantages over using smaller sample sizes, including the ability to conduct subgroup analyses, retain statistical power, and perform more complex analyses.

Overall, we do not believe synthetic data should completely replace humans – at least not in product testing. In the 1997 movie *"Good Will Hunting"*, the late actor, Robin Williams, portrayed a professor who mentors a young genius, played by Matt Damon. The prodigy holds a vast amount of knowledge, due to his superhuman ability to absorb information from books. In one of the scenes in the movie, the professor counsels the prodigy on the distinction between book knowledge and real-world experience. The professor says, "But I'll bet you can't tell me what it smells like in the Sistine Chapel. You've never actually stood there and looked up at that beautiful ceiling". True knowledge comes from living life, not from books, pictures, videos or any other representations of the real world.

Like the prodigy in the movie, an AI can be fed all the knowledge in the world that exists, but an AI will never be able to experience the world like a human being. There is something unique and beautiful about being human and being able to feel of the warmth of the sun on our face, enjoy the melody or beat of music, or the ability to behold a beautiful sunset – experiences that AI will never fully be able to replicate, no matter how advanced it becomes. How humans react to products, or life in general, is not captured solely in the brain as factual or semantic knowledge, our bodies and sensory experiences play a significant role, too.

> True knowledge comes from living life, not from books, pictures, videos or any other representations of the real world.

Key takeaways

01

Synthetic data will never be human. AI alone can never echo our product experiences, which combine the five senses, emotions, expectations, and context. Therefore, our goal is to augment human input with synthetic data, not replace it.

02

Accuracy hinges on the training data. The value of synthetic data is not binary (good or bad); the accuracy of synthetic data depends on many factors including the differences in the data we are trying to replicate, and the representativeness of the real-world data we are training an AI to learn from. The use of synthetic data should be strategic, considering the associated risks and benefits.

03

When accurate, it can power product testing. Synthetic data can boost market research agility, making it ideal for resource-intensive areas like product testing – reducing costs, saving time, with additional benefits for detailed sub-group analyses.

Endnotes

1 Guidi, M., Hubert, B., Sava, C., & Timpone, R. (2024). Synthetic Data: From Hype to Reality - a guide to responsible adoption. Ipsos POV

2 Illic, Maya, Bangia, Ajay, Legg Jim (2024). Conversations with AI Part V. Is there depth and empathy with AI twins? Ipsos Views

3 Moore Chris, Stronge Cameron, Bhudiya, Manjula (2024). Judgment Day: The Machines Have Arrived – But how good are they at answering choice experiments? Sawtooth Conference.

4 Yan Tao, Olga Viberg, Ryan S. Baker and René F. Kizilcec (2024). Cultural bias and cultural alignment of large language models. PNAS Nexus, Vol. 3, No. 9

5 Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate (2022). Out of One, Many: Using Language Models to Simulate Human Samples. arXiv.

6 AI-based Deep Learning is a way for computers to learn by analyzing large amounts of data and finding patterns, much like how humans learn from experience. It uses neural networks inspired by the human brain to recognize information and make decisions without being explicitly programmed.

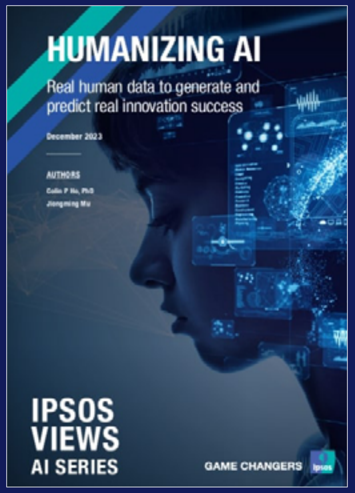
7 Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio (2014). Generative Adversarial Networks. arXiv.

8 Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.

9 Reynolds, N., Zach, J., Cho, J, Ho, C. (2021). Towards more agile and efficient product testing. Opportunities and limitations for smaller sample sizes, Ipsos POV.

10 We also compared the results from the all-synthetic data to the all-human data; the findings are like those reported in this paper

Further Reading


HUMANIZING AI
Real human data to generate and predict real innovation success
December 2023
IPSOS VIEWS
AI SERIES
GAME CHANGERS

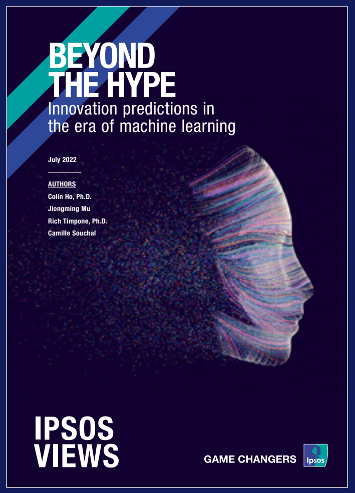
IPSOS VIEWS
SYNTHETIC DATA
From hype to reality – a guide to responsible adoption
Michel Guidi
Benoit Hubert
Ciprian Sava
Rich Timpone, Ph.D.

IPSOS VIEWS
THE NEW ERA OF INNOVATION
Shattering the Stage Gates with Generative AI
Dr. Nikolai Reynolds
Jiongming Mu

WE'RE MORE THAN OUR SENSES
Taking product development to the next level
September 2022
IPSOS VIEWS
GAME CHANGERS

TOWARDS MORE AGILE AND EFFICIENT PRODUCT TESTING
Opportunities and limitations for small sample sizes
ABSTRACT

EXPLORING THE CHANGING AI LANDSCAPE
From Analytical to Generative AI
April 2023
IPSOS VIEWS
GAME CHANGERS


BEYOND THE HYPE
Innovation predictions in the era of machine learning
July 2022
IPSOS VIEWS
GAME CHANGERS

CHALLENGING THE STATUS QUO
What makes a new product succeed
March 2022
IPSOS VIEWS
GAME CHANGERS

ADAPTING PRODUCT TESTING IN CHALLENGING TIMES
Be Contactless; Leverage Technology, Get Social
IPSOS VIEWS
GAME CHANGERS

JANUARY 2025

THE POWER OF PRODUCT TESTING WITH SYNTHETIC DATA

Humanizing AI series, part two

AUTHORS

Colin Ho, Ph.D
Chief Research Officer,
Innovation and Market Strategy
& Understanding, Ipsos

Dr. Nikolai Reynolds
Global Head of Product Testing,
Ipsos

The **IPSOS VIEWS** white papers are produced by the **Ipsos Knowledge Centre**.

www.ipsos.com
@Ipsos