B2B International
a dentsu company

The Role of Synthetic Data in B2B Market Research

An Introduction To Synthetic Data

Synthetic data has been posed as the solution to many of the challenges faced by market researchers. It is claimed that artificially-generated data can mimic real-world responses, significantly changing the outlook for traditional market research data collection methods.

It is appealing in theory because it would be possible to slash the cost of B2B research and create simulated data sets from any audience (including niche, hard-to-reach B2B profiles) in virtually no time at all. Headaches that B2B researchers face in recruiting and surveying targets would be a thing of the past; we would no longer need to worry about the practicalities of reaching people or debate what to do if we are not able to achieve a “robust” sample of B2B decision makers.

On the other hand, sceptics caution that the benefits of AI-generated data are overstated and outweighed by its potential drawbacks – put simply, this group argues that you are better off with smaller sample sizes with responses from real people rather than augmenting data sets with artificially-generated data.

As a specialist B2B agency, we wanted to put synthetic data to the test to objectively evaluate its potential applications in B2B market research today and in the future.

What we need to be mindful of in B2B market research

Given our area of experience within the market research category, our experiment focuses on B2B only. There are some specific nuances of B2B research that we must consider when discussing synthetic data.

Sample sizes in B2B research are generally much smaller than B2C – there is simply less data for AI models to learn from. The overall population of individuals who make decisions around specialty chemicals purchased by businesses is much lower than the population of consumers who purchase shampoo, for example. AI will generally have less primary data to draw from, therefore making it more difficult to create a robust model.

If any non-primary real-world data is being used to train the models (such as from online reports, public data repositories or data gathered from social media), it is much more challenging to find data relating to business decision making online. For example, there are thousands of people who post on social media their opinions towards political parties – but very few people who discuss their perceptions of engine transformer brands. It is therefore very difficult for any B2B AI models to take advantage of publicly available data as an additional training source.

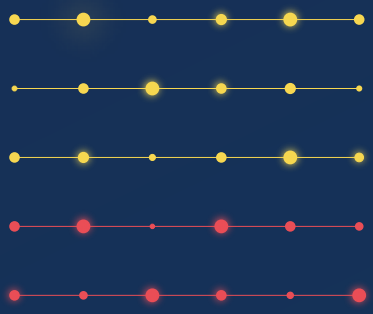
In addition to being smaller in number, business decision makers are harder to reach (for the purposes of research) than consumers. B2B research projects are therefore usually commissioned to inform significant strategic decisions - ensuring that the right questions have been asked to the right people is at the heart of all good market research and this requires both time and investment.

Decision making behaviors within organizations are complex. Our Superpowers Index research demonstrates that the decision making unit is growing and that the buying journey is getting longer. Purchasing decisions are also less homogenous in B2B scenarios when compared to B2C. This makes the drawing of inferences from adjacent decision-making behaviors less reliable. You might be able to learn some things about how people choose an ice-cream from how they choose a chocolate bar, or even a film, but these sorts of cross-references are less likely to hold up when comparing building aggregate purchases with selecting a new partner for cloud security services.

How synthetic data can be used in B2B market research

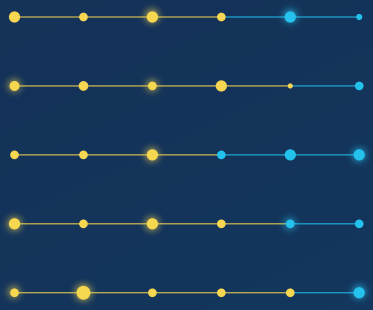
There are several potential use cases for synthetic data in B2B market research. The most interesting to us right now are applications around increasing sample sizes (i.e., creating more respondents) and how to expand data sets without increasing survey length or fieldwork (i.e., asking more questions).

Potential applications of synthetic data in B2B market research:



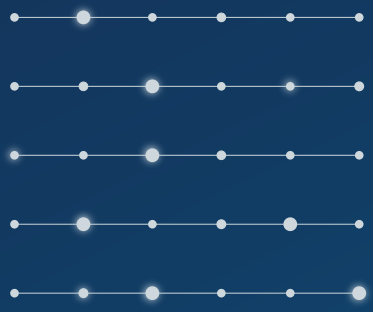
Creating more respondents:

To augment existing data sets, theoretically useful in situations where it has not been possible to reach a robust number of interviews during fieldwork or where audiences are niche, and it would not be possible to create a large data set.



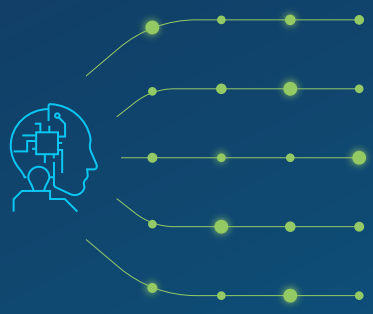
Asking more questions:

Creating answers to questions that were not asked to respondents as part of the survey. This can be helpful to complete surveys where there are high drop-out rates or where questions have been added to the study part-way through.



Improving privacy and compliance:

Masking sensitive information in a real data set to remove any personal data included in the data set. Useful to comply with privacy regulations such as GDPR and CCPA.



Training AI language models:

Synthetic data can be used to train AI language models to answer questions as if they were a member of the population you are looking to study. This can allow for greater exploration into data sets, without having to exact real-world data, as well as testing out new ideas with a target audience without risking any details leaking to the competition.

Putting synthetic data to the test in B2B

What we did:

We set out to put synthetic data to the test with a large, robust B2B data set. As part of our annual Superpowers Index – the world's largest systematic study of B2B buying behavior globally – we survey more than 3,000 individuals to understand how decisions are made in their organizations. The study is global and covers 4 key purchase categories: Manufactured goods, financial services, professional services, and tech. Since 2021 there has been a level of consistency in the questionnaire so we can track perceptions over time, as well as some new questions on emerging topics of interest – for example we'll be exploring the impact of AI and tariffs in the 2025 wave of the research.

This data set was perfect for the trial, as we knew what the complete data set looked like over multiple waves of the research. We tested synthetic data in two ways...

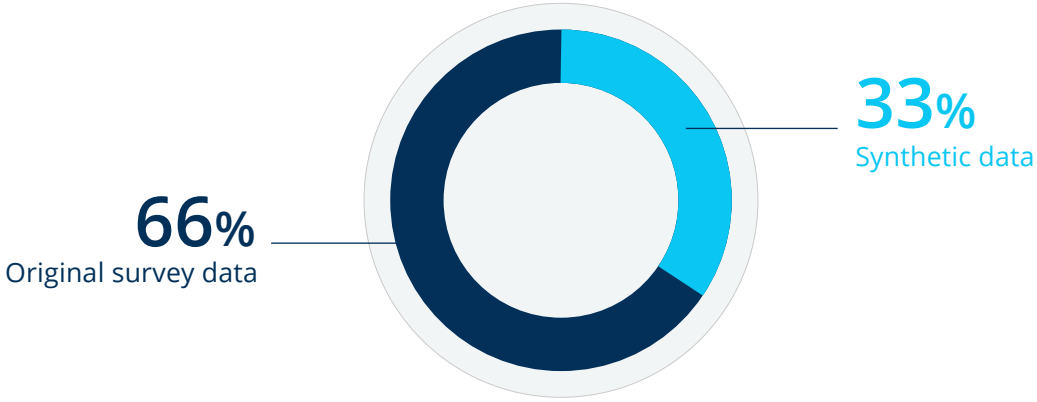
Experiment 1: Augmenting the existing data set

We provided a supplier of synthetic data with a random subset of 2/3 of the sample from our 2024 survey data and asked them to simulate the final third.

The question we wanted to answer was simple and relevant to B2B research:

"If we are not able to reach a representative sample of decision makers, does augmenting the existing data give a better or worse result than doing nothing?"

The augmented data set was made up of:



We analyzed the augmented data set by looking at the average accuracy of the responses and the extent to which the data set exhibited the same patterns of significant differences. To keep the task manageable, we looked at the overall pattern as well as three cross-breaks that have been useful to develop the story and narrative around the research findings – company size, seniority of the respondent, and whether the business was a pure-play B2B business or also sold to consumers.

To evaluate how well the augmented data set performed we looked at three areas of the survey, each with different objectives. These were: decision making drivers, green targets set by organizations, and the communication channels used – a nice mix.

What we found:

Across the four key areas we explored, the overall data set and the three cross-breaks (company size, seniority, and type of business), we looked at the accuracy of the augmented data set vs. the full data set. In the table on the right, this is the **"augmented data"**.

For the second scenario we looked at the accuracy of the partial data set vs. the full data set – this is where we can see what the implications of **"doing nothing"** would be, i.e. what would the implications of closing fieldwork early be on the survey results?

Overall	Augmented data	Very high accuracy
	Do nothing	Very high accuracy
Company size	Augmented data	Acceptable accuracy
	Do nothing	Acceptable accuracy
Seniority	Augmented data	Acceptable accuracy
	Do nothing	Good accuracy
Type of business	Augmented data	Low accuracy
	Do nothing	Good accuracy

First, we'll look at **accuracy** at an overall level: the augmented data set mirrored the full 2024 data set. However, this was not the case when we started to dig into key cross breaks – there were significant variances when we drilled down by company size, seniority, and type of business. Even at an overall level where the synthetic data achieved a high level of accuracy, it was no more accurate than what could be achieved by doing nothing, i.e. closing fieldwork early and working with a smaller sample size. This is because the reduced sample error that comes from filling out our sample numbers is more than offset by the modelling errors introduced by the addition of synthetic responses.

Secondly, the synthetic data exhibited a smoothing effect, reducing the **variance** in the data – potentially masking real-world fluctuations that happen in data sets, particularly bearing in mind what we said earlier about decision making behaviors being less homogenous in B2B.

The final part of our test relates to a **regression model** we run on the Superpowers data set each year to understand the most influential B2B decision drivers. We were keen to see the impact of using the augmented data set for this analysis. There was very little similarity between the model that was created with the augmented data set; in fact running the model on the partial data (66% of the survey completes) resulted in much greater similarities.

Experiment 2: Simulating new waves of data

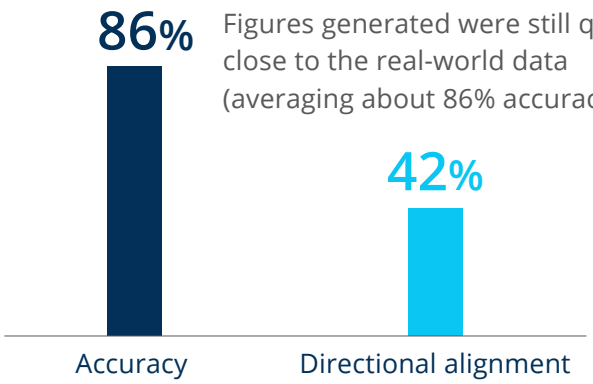
The second question we wanted to answer was:

"How accurately can we simulate new waves of the research?"

In this scenario, we provided data from 2023 and asked the provider to replicate the survey data from 2024. The benefit of this being that we know what the real 2024 data set looked like so we can compare.

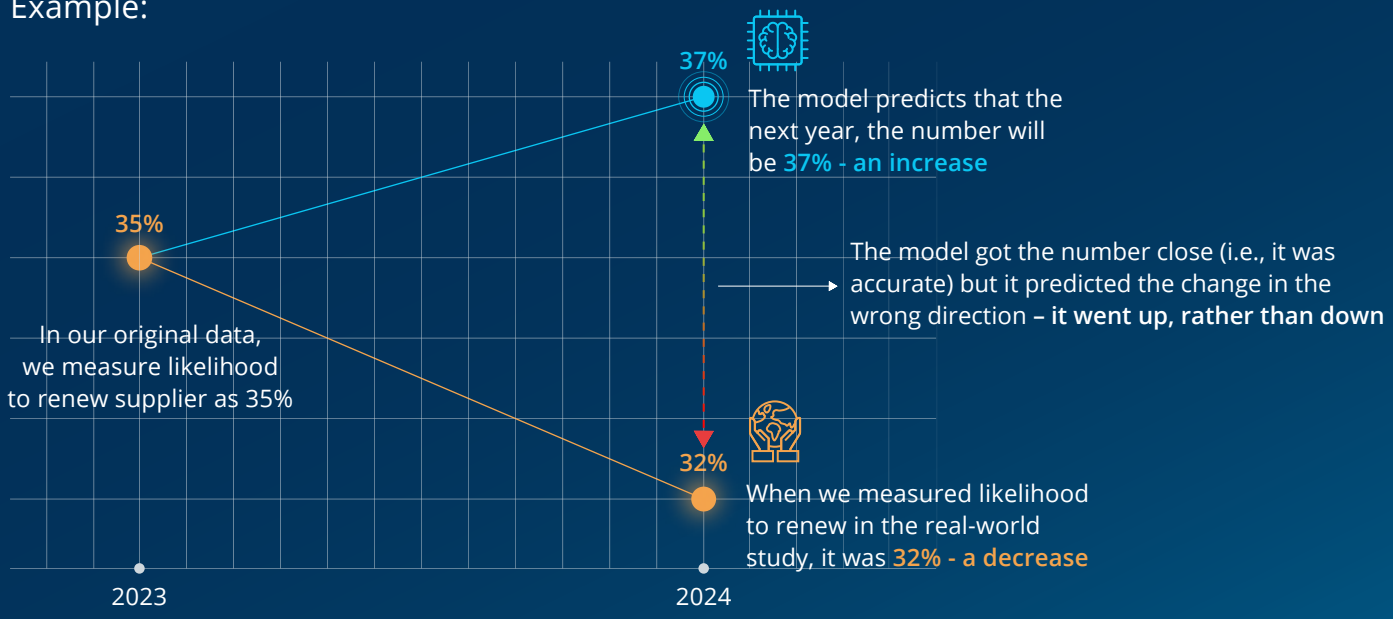
To assess the performance of the synthetic data, we looked at the average accuracy of the responses and the extent to which the data sets exhibited the same directional changes from one year to the next.

What we found:



However, when we look at directional alignment – whether the synthetic data correctly captures the trend over time, only 42% aligned with the expected trends. This is worse than just guessing – if you were to guess you'd expect to get the direction right 50% of the time, like flipping a coin.

Example:



The synthetic data clearly struggled to predict changes over time. This is a clear limitation on its suitability for longitudinal studies (e.g., brand tracking, U&A tracking, CX tracking) unless it is paired with real-world data gathered in the correct time period.

Overall Thoughts

In summary, at an overall level, augmenting real-world data with synthetic responses does not offer significant benefits over standard fieldwork (at least in the scenarios we've tested). At this point in time, the results suggest that it is better to do nothing and work with a smaller data set than augment with synthetic data in a bid to increase the amount of data that there is to work with. Similarly, attempting to predict future waves of data based on already collected sample is not advisable (you might be better off just assuming no change).

That said, this is only one potential application for synthetic data in B2B and technological advancements are expected to enhance accuracy and reliability over time. We expect that things will change quickly when it comes to synthetic data and these sorts of use cases may soon be more viable.

In addition, things already look more promising when looking at filling in missing answers in partially completed survey responses. This may allow for shortened survey lengths with each respondent only completing a subset of the full questionnaire, cutting costs and increasing incidence rates.

The training of language models to act as simulated qual interview respondents also offers an exciting first step in gauging high-level reactions to potential new offerings as well as mapping out important topics to cover in any quant studies that might follow.

It's clear that AI is extending our reach and already helping us do more, faster. Choosing where and when to adopt each emerging application is going to be the key to successfully harnessing this new technology.

B2B International

a dentsu company

To discuss how our tailored insights programs can help solve your specific business challenges, get in touch and one of the team will be happy to help.

www.b2binternational.com