QUEST MINDSHARE™

The Real Potential of Synthetic Data in Market Research

FAQ from Quest's presentations, webinars, and direct investigations

Following Quest's several presentations, conference talks, and webinars about synthetic data modeling, we received a remarkable number of insightful questions. We've compiled and answered the top 32—ranging from model mechanics to ethical use cases. Whether you're just exploring synthetic data or you're hands-on with modeling, there's something here for you.

QUEST MINDSHARE™

? Can synthetic data modeling answer a question that you have not directly asked to the sample?

Short answer: No.
Think of it as very advanced weighting, using different techniques but achieving the same overall result of building on existing data to show it in a new way. Algorithms are trained using actual data—they can extend and simulate relationships within that data, but they can't generate answers to questions never asked in the first place. Think of it as advanced weighting, not mind-reading. Real respondent input remains essential.

? How does the modeling handle representation and potential bias?

Modeling can't compensate for what's missing. It can only reflect the data it's given. If a group is entirely absent, no model can recreate it. Models work from patterns in known variables like demographics or behavior.

? Can you share more real-life use cases?

Absolutely. In our presentations, we're limited by time for what we can show. In work previously presented by Scott Worthge and Kyle Hope, such as at both ESOMAR World Congress 2024 and ESOMAR LATAM 2025, we show detailed results of our direct testing of synthetic data modeling. These presentations go deeper into use cases and model performance metrics. If you're interested, we're happy to share the full decks.

? Can synthetic data be used outside of surveys?

Absolutely. It's widely used in fields like autonomous vehicle simulations, manufacturing, and fraud detection. In research, it's especially helpful in augmenting data, modeling behavior, and simulating marketing scenarios.

? Does synthetic modeling work as well in B2B?

While no published B2B-specific studies exist yet, the math behind the modeling doesn't differentiate between B2B and consumer data. Quest is actively testing B2B applications with results expected later this year.

? When would it make sense to use fully AI-generated (synthetic) personas?

These can be useful for validating ideas quickly based on known data. However, they're limited to their training and may miss nuance. Purely synthetic audiences are best used directionally, not as substitutes for real-world respondents.

? Are there clear use cases where NOT to use synthetic data/audiences?

Yes—particularly with 100% qualitative data, extremely small base sizes, or when interactive/subjective content is involved. As a rule of thumb: if weighting isn't appropriate, synthetic modeling likely isn't either.

? Are there any reliable academic or industry sources on this?

Both Fairgen and Livepanel publish white papers detailing their technology. Few truly independent studies exist, but that's changing. We'll present more results later this year at ESOMAR NA and other conferences.

?Does synthetic modeling preserve inter-variable relationships?

Yes, this is one of its strengths. The process preserves correlations between questions by simulating responses based on overall dataset structure—not just one variable at a time.

?Isn’t synthetic data just perfectly correlated with training data?

Not exactly. Although it’s derived from existing data, the model reconfigures known information to build new respondent profiles. Multiple models are tested, so even users with identical characteristics may receive different imputed responses.

?Doesn’t using biased training data just amplify the bias?

It can. That’s why we stress validating your starting data. We’ve been testing this by comparing synthetic to real survey responses. If your original data is biased, your modeled data will be too—just like with weighting.

?Can clients just send in data and ask for modeling?

Not quite. Modeling depends on having training data from the same audience/context. You can’t model a U.S. audience from European data, for example. Inputs must match the population you're modeling.

QUEST
MINDSHARE™

?Can synthetic data create echo chambers if used repeatedly?

Yes—especially in trackers. Without regularly updating real input data, you risk compounding assumptions. Best practice is to retrain models with fresh data periodically.

?How can we trust this data if risks can't be fully mitigated?

We test, iterate, and compare. For many clients, modeling starts as a parallel process—run side-by-side with traditional data collection—to assess reliability before scaling up.

?Does modeling one question at a time cause overfitting?

It can, but our process mitigates this by running multiple smaller models for each question, comparing their outputs to choose the best fit. It's like cross-validation to reduce noise.

?Is there a minimum number of real completes needed?

Generally, start with 50 per key subgroup or at least 50% of your overall modeling target. The more detailed the profiling (e.g., demographics, psychographics), the fewer completes you'll need.

?How is this different from traditional weighting?

Think of the two processes as if you were trying to expand and change a picture – weighting is like stretching pixels, while synthetic modeling redraws the image in high definition.

? Are there any specific demographic, sociographic or other variables that tend to be more critical in the accurate modeling and imputing of data? If so, what are the key variables that need to be incorporated ?

Not enough formal testing exists yet. But more inputs—especially demographics and behavioral traits —lead to more robust modeling. Psychographics are promising but still exploratory.

? Can you compare this to imputation or bootstrapping?

Yes. While mean or basic imputation ignores relationships between variables, synthetic modeling simulates multi-dimensional answers. It's more dynamic and usually more accurate.

? How does it compare to methods like MICE or SMOTE?

We're exploring this now—comparing various imputation strategies on identical datasets. Early indicators suggest synthetic modeling outperforms for multivariate preservation.

? What platforms or providers do you use?

We've worked closely with Fairgen and Livepanel. Both offer proprietary algorithms and have engaged openly with us about their methodologies.

? Isn't synthetic data more useful for quick, directional insight than hard decisions?

We agree. For now, synthetic data is best used to augment, validate, or accelerate —not to entirely replace—research in high-stakes decision-making contexts.

QUEST MINDSHARE™

? Do clients really want this yet?

Client demand is catching up. Providers have been building capabilities for years, and those experimenting now are positioning themselves ahead of the curve.

? Is this being used in qualitative research too?

Others are testing this (e.g., Ipsos, ToLuna), but Quest's focus is on quant applications. That said, qualitative personas are an exciting frontier. However, we're already seeing some integrations between personas and modeling emerging, a "qual-quant" approach as we'd call it.

? How does this relate to AI democratizing insights?

AI is empowering more people to explore data—but it must be used responsibly. Modeling is one tool in an expanding toolkit that researchers have not had access to before a year or two ago. Despite the development of algorithms and the fundamental basis of modeling data being very well established, the actual implementation of synthetic data modeling for market research projects is very much a developing frontier.

? Is it useful for low-incidence audiences?

Yes. We've seen solid results modeling hard-to-reach segments—like deck builders or niche professionals—when enough real seed data is available.

? How much variance is there between synthetic and real data?

In our testing, very little—especially for attitudinal scale data. Variance increases when real data has high standard deviation (i.e., spread out responses), which makes modeling less predictable. Quest is actively testing modeling applied to several types of data and will be publishing results in 2025-2026.

QUEST MINDSHARE™

? Can you model a very specific subgroup with limited overlap?

Only if you have 50+ real completes with most of the traits you're targeting. We're actively testing whether overlapping partial traits can be modeled effectively.

? How does synthetic modeling compare to mean imputation?

It's significantly more accurate. Mean imputation flattens relationships; synthetic data modeling preserves them through multidimensional analysis.

? What tool generated your synthetic data?

Fairgen and Livepanel. We're always looking for new providers to test, but they've been the most open and transparent in explaining their processes.

? What about using AI panels built from LinkedIn bios or CustomGPTs?

That's cutting edge and fascinating—but very hard to validate. Real-world side-by-side testing is the only way we're comfortable trusting those results.

? Is synthetic data the same as AI?

Not quite. AI (like GPT) can generate data, but synthetic data modeling starts from real, structured input and builds from there. Garbage in, garbage modeled. Data quality is still king.

QUEST
MINDSHARE™

Have more questions?
If you want to explore any topic in more detail, contact the Quest Mindshare team. We're committed to transparency and sharing our findings as we explore what synthetic data can unlock for our industry.

Reach out for more to:

Kyle Hope
Diretor of Supply & Partnerships
khope@questmindshare.com

Scott Worthge
Research Director
sworthge@questmindshare.com