AUGUST 2026

State of Synthetic Data

Governance + The ECHO Index

ksrinc.com

KS&R

DOWNLOAD THE DECK

Today’s Speaker

Ben Cortese, PhD

Decision Sciences & Innovation

Senior Vice President

Here's where we're headed with today's agenda.

01 **Synthetic Data**
What we mean by synthetic data and its applications in insights

02 **Synthetic Governance**
KS&R's ECHO Index and how it applies to synthetic data

03 **The ECHO Index in Practice**
Measuring the fidelity of synthetic data and a roadmap to model improvement

04 **Practical Guidance**
Trends we're seeing from governance and evaluation

First, let's align on what we mean by synthetic data.

Synthetic data covers a broad spectrum of definitions and use cases.

Methodology	Description	Requires Gen AI	Use Cases	Confidence
Extrapolation	Extend a statistical model beyond the input data to predict a plausible outcome.	**No**	○ **Extend price curves** beyond the price range tested ○ **Predict future wave results** from prior wave	**Low** No guarantee that trends will hold outside of a model's context window.
Imputation	Statistical, machine learning, and/or deep learning algorithms fill in values for missing data.	**No**	○ **Reduce survey length** by intentionally omitting questions via a statistical design. ○ **Complete Don't Know / NA responses** (Note: while this is common practice, it is a grey area from a statistical viewpoint)	**High** Methods in this space are well established and accepted.
Sample Boosting	Synthetic data augmentation via models trained on real world reference data (human-based) and mimics the original data structure. *Cannot extend results outside of training data.*	**No**	○ **Reduce the overall sample size** needed via synthetic respondents (generate new rows) ○ **Bolster sample sizes for hard-to-reach audiences**, enabling reporting for small segments	**Moderate** Validation and governance via the ECHO Index quantifies confidence.
Digital Twins	Individual models that require a 1:1 match to a human, process, system, etc., tuned to respond and behave like the data they are trained on.	**Yes**	○ **Extend questionnaire to adjacent topics** not included in the original research (generate new columns) ○ **Enables synthetic "recontact"** without the need to re-field ○ **Have drill down conversations** to understand the "why" behind decisions	**Low** There is no validation and governance mechanism to date. Directional at best.
Synthetic Personas	Aggregate models trained to simulate a broader target audience or population. This can be viewed as a generalization of digital twins.	**Yes**	○ **Create interactive personas** from a segmentation study ○ **Enable a natural language engagement model,** extending insights and enabling quick-turn follow-ups. ○ **Rapid testing of messages, concepts, etc.** to trim long lists ○ **Ideation and early stage planning**	**Low** There is no validation and governance. Directional at best.

Visualizing tabular synthetic data reinforces where we have higher confidence in synthetic applications.

Real World Data (RWD)

01

Respondent ID	Var 1	Var 2	Var 3
1	X	Y	Z
2	X	Y	Z
3	X	Y	Z

03

Synth Var 101	Synth Var 102
α	β
α	β
α	β

Digital Twins

Generate Additional Columns

- AI-derived variables build new columns, enriching the data
- Provides insights to new questions
- Requires LLMs

02

Synth 1	x	y	z
Synth 2	x	y	z

04

??	??
??	??

Sample Boosting

Generate Additional Rows

- Synthesize net-new responses to questions included in the training data
- Generally viewed as most reliable
- Does not require LLMs

Synthetic Audiences

Generate Both

- Extends new column generation into a space where validation is ill-defined.

KS&R's ECHO Index was created to evaluate the viability of sample boosting.

Why start with sample boosting?

- This is the most mature methodology in use today and has trusted tools for evaluation.
- Sample boosting has the potential to solve a variety of existing challenges including reduced data collection costs, less time in the field, and broader access to hard to reach audiences.

What about digital twins?

- Digital twins create new data with no real world reference, presenting a unique evaluation challenge.
- The ECHO Index will extend to digital twins as part of our synthetic data governance roadmap.

How does ECHO Index scoring work?

The ECHO Index measures how close synthetic data is to the baseline distribution.

It generates confidence in synthetic models via a rigorous, transparent evaluation framework.

There are **two key components** to evaluation:

01

Statistical

Are univariate and bivariate relationships preserved?

02

Machine Learning

Can a classifier tell real from synthetic data?

The **ECHO Index** applies to any standalone data set.

The real world data is used to establish a baseline **ECHO Index** for scoring comparisons.

Synthetic data is scored via the **ECHO Index**.

The **ECHO Index** for synthetic data is compared to the baseline, quantifying confidence.

Full details were unveiled at the **Sawtooth Research Conference (SRC)** in Berlin in early May, 2026.

Statistical scoring covers common question types using multiple evaluation criteria.

Question Type	Sample Question Text
Likert Scale	On a scale of 1 to 7 with 7 being the most, how willing are you to recommend Brand X to a friend or family member?
Single Choice	Considering the following brands, which do purchase from the most frequently? (Select one)
Multi Select	Considering the following brands, which have you heard of? (Select all that apply)
Open-end	Explain why you purchase from this brand most often.
Rank	Rank the following brands from most expensive to least.
Misc (e.g., slider, etc.)	

The most common question types are treated as categorical, numeric, or binary which align well with systematic statistical testing.

Unstructured and less common question types will be handled in future iterations of our evaluation.

Distribution Preservation

Ensemble scoring, comparing real and synthetic data from several angles will be used to generate the distribution score for each question, leveraging:

- Kullback-Leibler (KL) Divergence
- Jensen-Shannon Divergence
- Total Variation Distance

Relationship & Dependency

Like to like variable comparisons measured through differences define relationship & dependency performance:

- Pearson correlation (numeric)
- Cramer's V (categorical)
- Phi coefficient (binary)

The Machine Learning Score measures how difficult it is for classification models to separate real from synthetic data.

Classification Methods:

Random Forest

XGBoost

Logistic Regression

- Apply to data sets with mixed variable types
- Scalable with minimal hyperparameter tuning
- Blend of historic and modern classifiers

Success Metrics:

ROC AUC

F1

Balanced Accuracy

- Applicable to all classifiers
- Balance accuracy and bias estimation

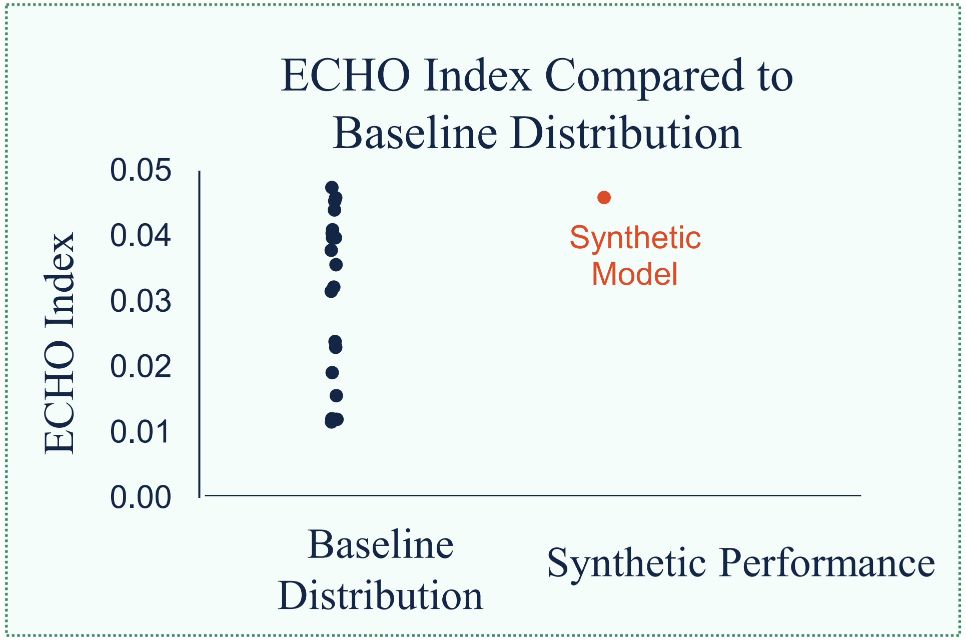
Cross Validation:

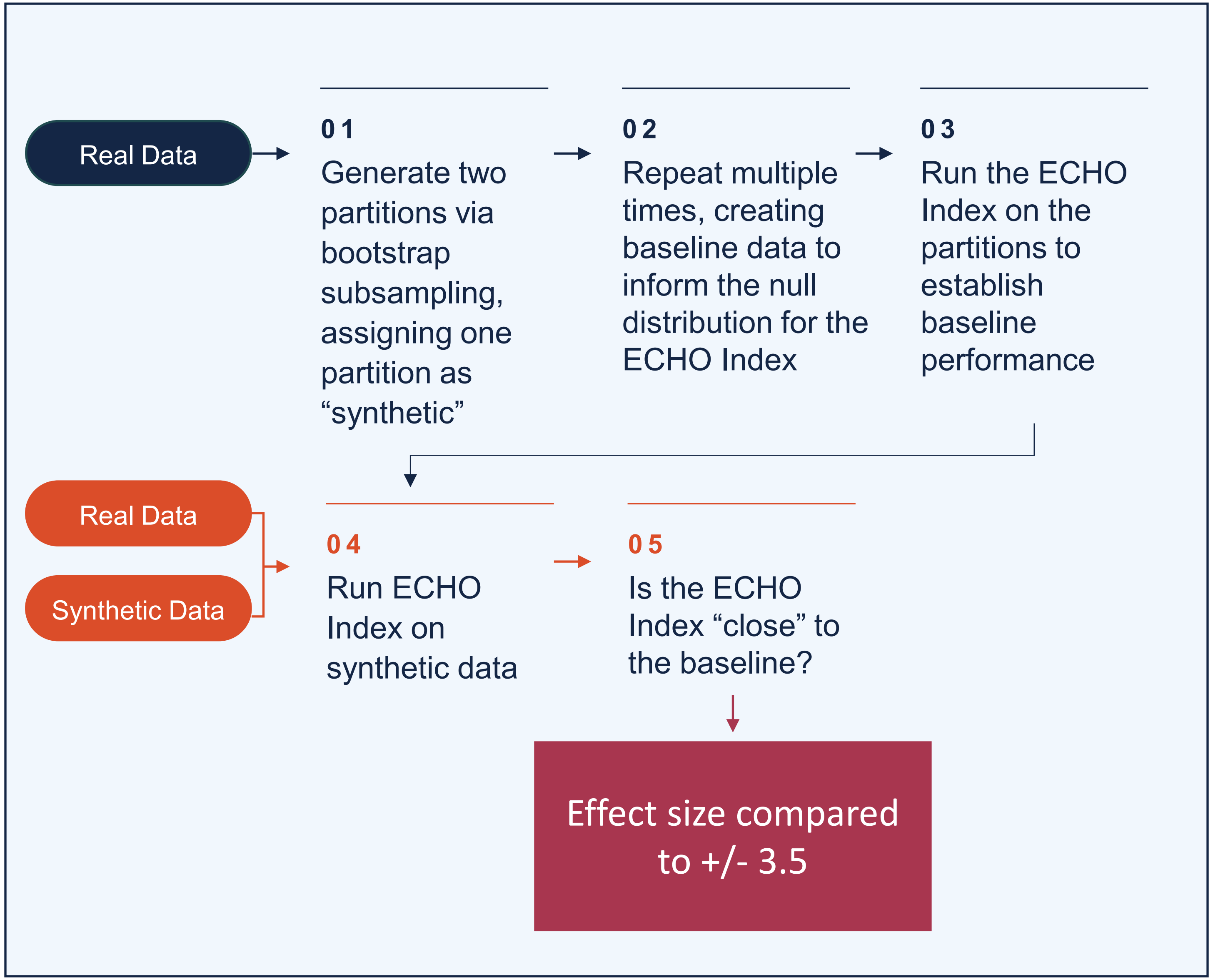
10-fold

- Average of success metrics taken across folds
- No true holdout set included as this doesn't apply to our core use case of synthetic data

Higher success metric scores indicate lower quality synthetic data, contrary to the traditional interpretation of classifier performance.

The full ECHO Index pipeline measures how close synthetic data scoring is to the baseline distribution.





The focus of the ECHO Index is statistical fidelity, a critical component of synthetic evaluation.

WHAT IS

statistical fidelity?

Fidelity measures **how well synthetic data preserves the statistical properties present in the real world data**. This includes the distribution or shape, relationships among other variables, and absence of synthetic artifacts.

WHY IS

fidelity critical?

Fidelity ensures the synthetic data is indistinguishable from the real world data it was trained on. **Evaluation through statistical and machine learning methods provides a roadmap to enhance synthetic models**, something that utility alone cannot inform.

WHAT ABOUT

utility?

Utility determines if downstream analytics and reporting would lead to the same result as human sample. This is challenging to measure directly and does not provide the “why” driving any misalignment.

Applying the ECHO Index shows where synthetic is fit for purpose

DOWNLOAD THE DECK

We evaluated several synthetic providers across two publicly available datasets.

THE PARTICIPANTS & MODELS

Synthetic Providers

- Provider A – Statistical/ML models
- Provider B – Statistical/ML models
- Provider C – Statistical/ML models
- Provider D – LLMs
- Provider E – LLMs

THE DATA

Paxton and Yang[1]

- Straightforward study with 11 questions **intended to give synthetic models the best chance of success.**
- Comprised of 518 human responses from a panel aggregator.
- US Consumer audience with focus on trust in tech

Twin-2K-500[2]

- Highly complex study with over 500 questions **intended to push the limits and challenge synthetic models.**
- Developed by researchers at Columbia University, including 2,058 respondents.
- A broad range of topics are covered, including demographic, psychological, economic, personality, and cognitive measures.

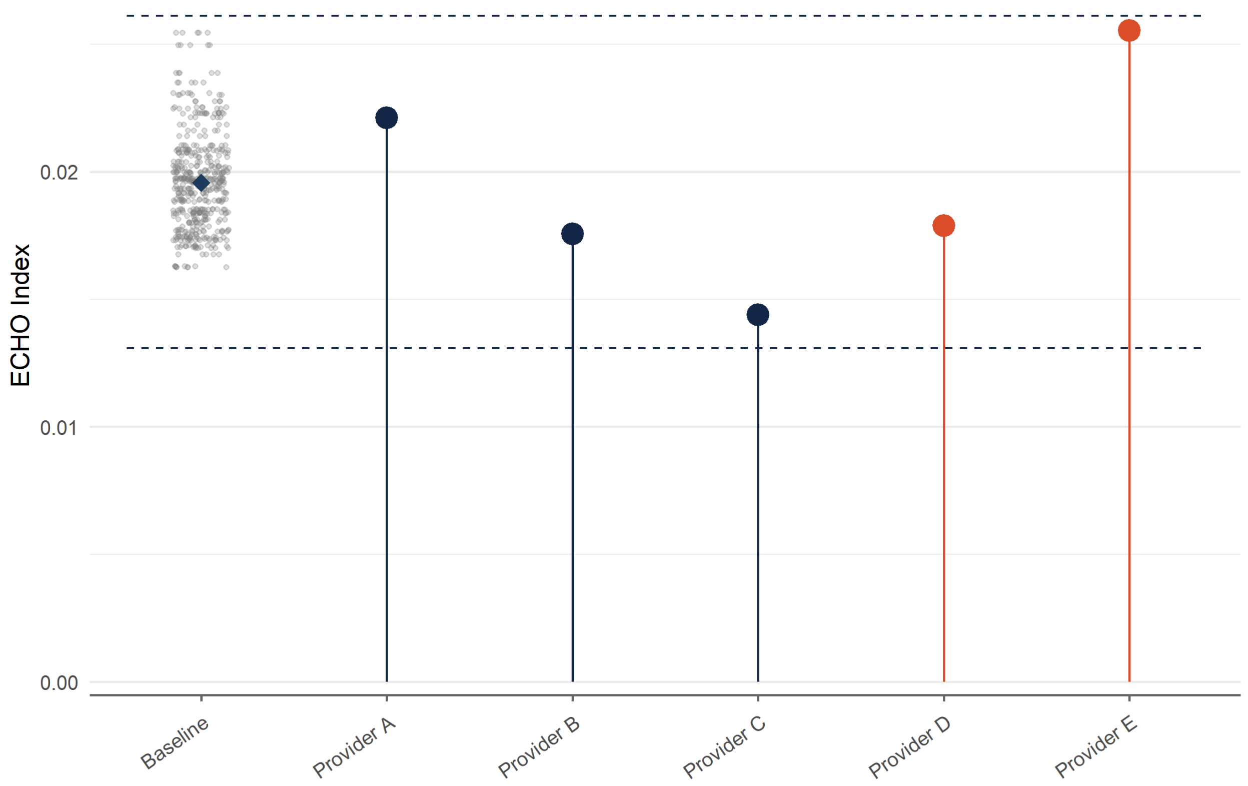
> Each model was used to create a synthetic copy of the data sets, doubling the size of the real data.

[1]Paxton, J., & Yang, Y. (2024, June 12). *Do LLMs simulate human attitudes about technology products?* [Conference presentation]. Quant UX Con.

[2]Toubia, O., Gui, G. Z., Peng, T., Merlau, D. J., Li, A., & Chen, H. (2025). *Twin-2K-500: A dataset for building digital twins of over 2,000 people based on their answers to over 500 questions*. arXiv. https://doi.org/10.48550/arXiv.2505.17479

Evaluation of synthetic data for Paxton and Yang demonstrates high fidelity synthetic data.

Results are promising, producing evidence that sample boosting is viable.



Provider	Approach	ECHO Effect Size Index
Provider A	Stats/ML	1.4
Provider B	Stats/ML	-1.1
Provider C	Stats/ML	-2.8
Provider D	LLM	-0.9
Provider E	LLM	3.2

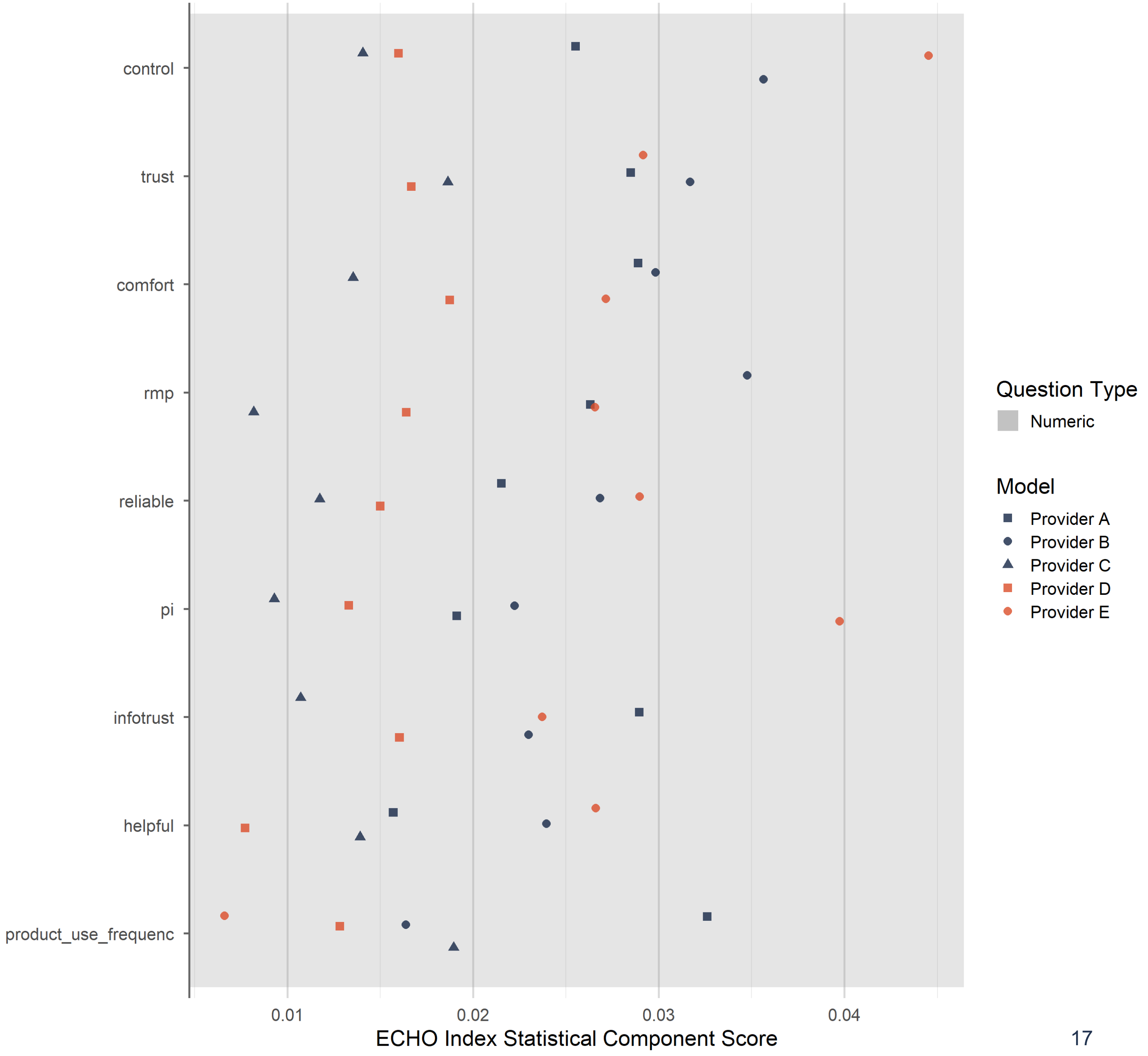
All synthetic providers produced ECHO Index scores within the fit for purpose threshold (+/-3.5 effect size), a positive sign for sample boosting.

Digging deeper into individual question performance reinforces synthetic model fidelity.

Random scattering of question-level statistical scores across providers demonstrates how both distribution and relationships among other numeric variables are preserved via these synthetic models.

There is no clear pattern of LLM vs. traditional synthetic model performance by question.

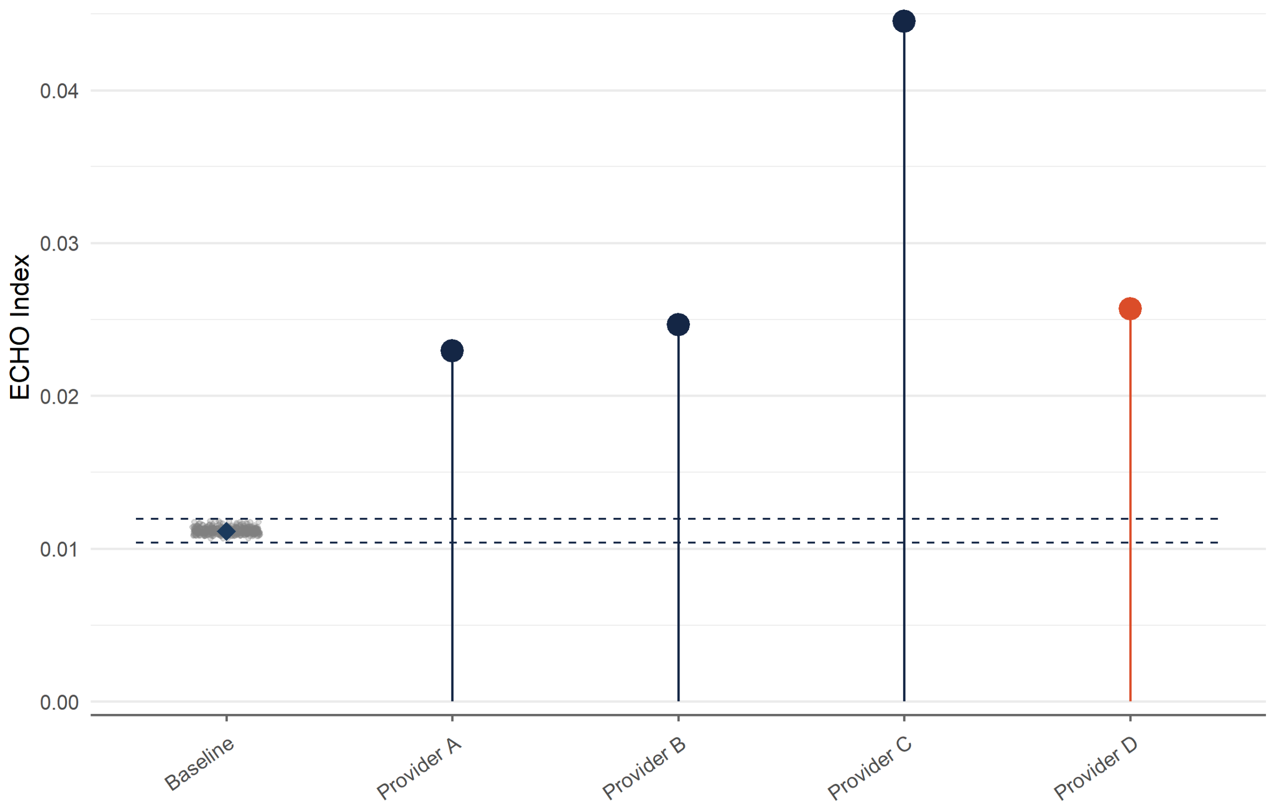
These are indicators that straightforward, simple studies are low hanging fruit for sample boosting.



KS&R

Synthetic data using the Twin-2K-500 study suffers from low fidelity across the board.

The complex nature of this study has gone beyond the limits of sample boosting capabilities.



Provider	Approach	ECHO Effect Size Index
Provider A	Stats/ML	53.0
Provider B	Stats/ML	60.6
Provider C	Stats/ML	75.0
Provider D	LLM	65.3

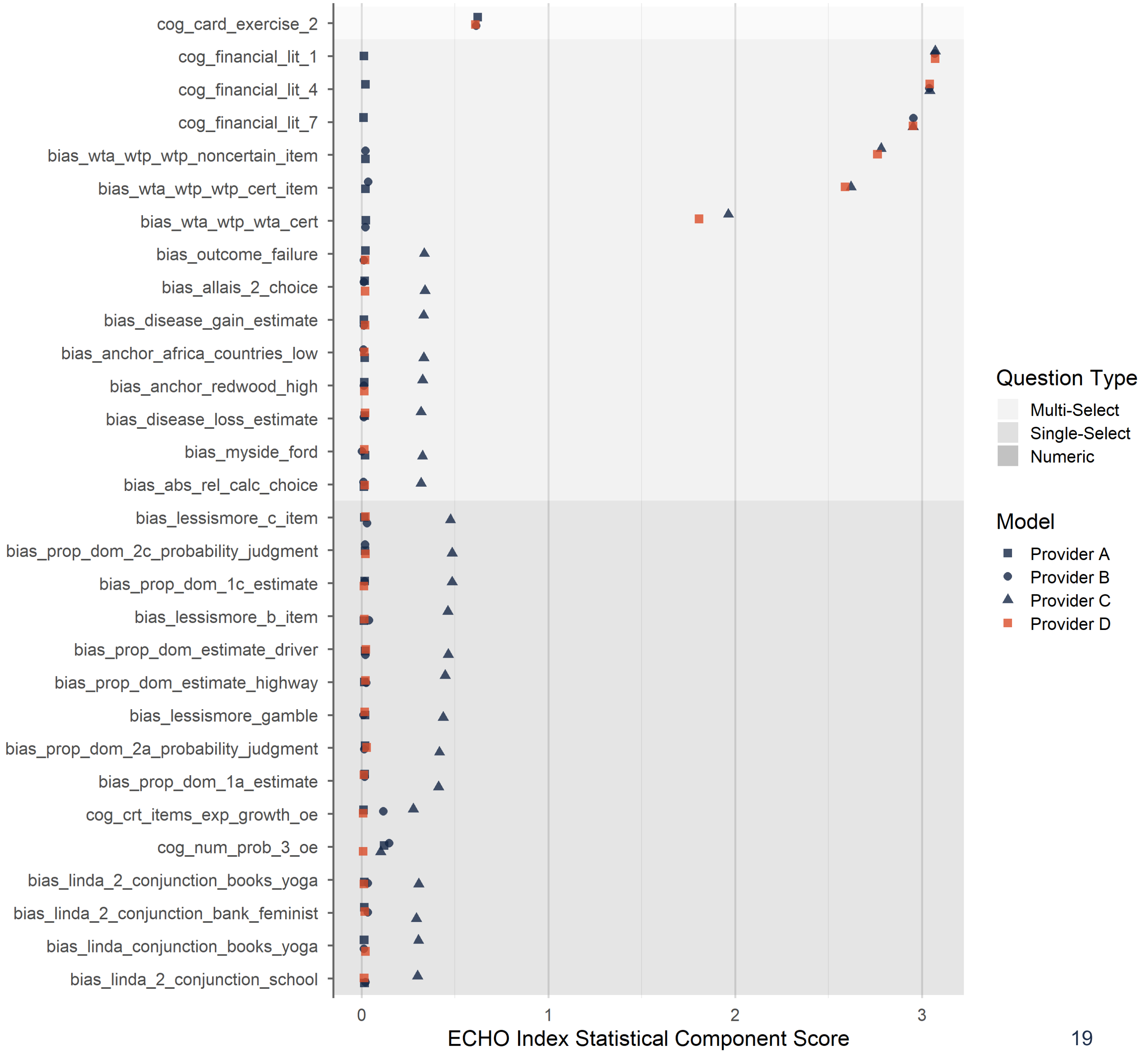
Provider performance is surprisingly similar, with all effect sizes at least an order of magnitude higher than the fit for purpose threshold.

This begs the question “what is driving poor model fidelity?”

Individual question performance showcases the primary drivers of low overall fidelity.

These results begin to build a roadmap for synthetic model iteration and enhancement.

> Similar to Paxton and Yang, we see no obvious pattern of LLM vs. traditional synthetic model performance by question, but there is nuance.
>
> This details the roadmap for either adjusting synthetic models or removing poor performing questions from the synthetic dataset.



KS&R

Deep dive into questions that break synthetic models

DOWNLOAD THE DECK

The One Question that Fooled them All

The cognitive card exercise had low fidelity across all models. This is likely due to erratic training data that is difficult for any synthetic algorithm to predict.

Do you know the correct answer?

Imagine you see a number of cards from a set of cards. Each card in the set has a capital letter on one side and a number on the other. Naturally, only one side is visible in each case.

For the set of cards, a rule has been stated: If there is an A on the letter side of the card, then there is a 3 on the number side. Four cards were drawn.

Below is the information visible for each card (letter or number).

Which of the following card(s) would have to be turned over in order to test the truth or falsity of the rule, assuming the cards shown are “A”, “F”, “3”, and “7”?

Financial literacy and valuation of mortality risk favor some statistical / ML models

At least one statistical / ML model demonstrated high fidelity, while the LLM failed to produce realistic looking data across the board.

Financial Literacy

Do you think that the following statement is true or false?

- A 15-year mortgage typically requires higher monthly payments than a 30-year mortgage, but the total interest paid over the life of the loan will be less.
- If you were to invest $1,000 in a stock mutual fund, it would be possible to have less than $1,000 when you withdraw your money.
- After age 70 and a half, you have to withdraw at least some money from your 401(k) plan or IRA.

Valuation of Mortality (Representative Example)

Imagine that when you went to the movies last week, you were inadvertently exposed to a rare and fatal virus. The possibility of actually contracting the disease is 4 in 1,000, but once you have the illness there is no known cure. On the other hand, you can, readily and now, be given an injection that reduces the possibility of contracting the disease to 3 in 1,000. Unfortunately, these injections are only available in very small quantities and are sold to the highest bidder. What is the highest price you would be prepared to pay for such an injection?

Mathematical reasoning doesn’t appear to fool LLMs

Both cognitive reasoning questions had strong performance from the LLM, while traditional methods struggled.

LLM Only

If the chance of getting a disease is 20 out of 100, this would be the same as having a [blank]% chance of getting the disease. Enter a percentage from 0 to 100.

LLM and One Statistical / ML Model

In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? Enter a number of days.

Practical guidance from synthetic projects

This exercise demonstrates the potential for synthetic data.

There is no “rule of thumb” for when synthetic works, case by case evaluation is essential.

Simple, straightforward data

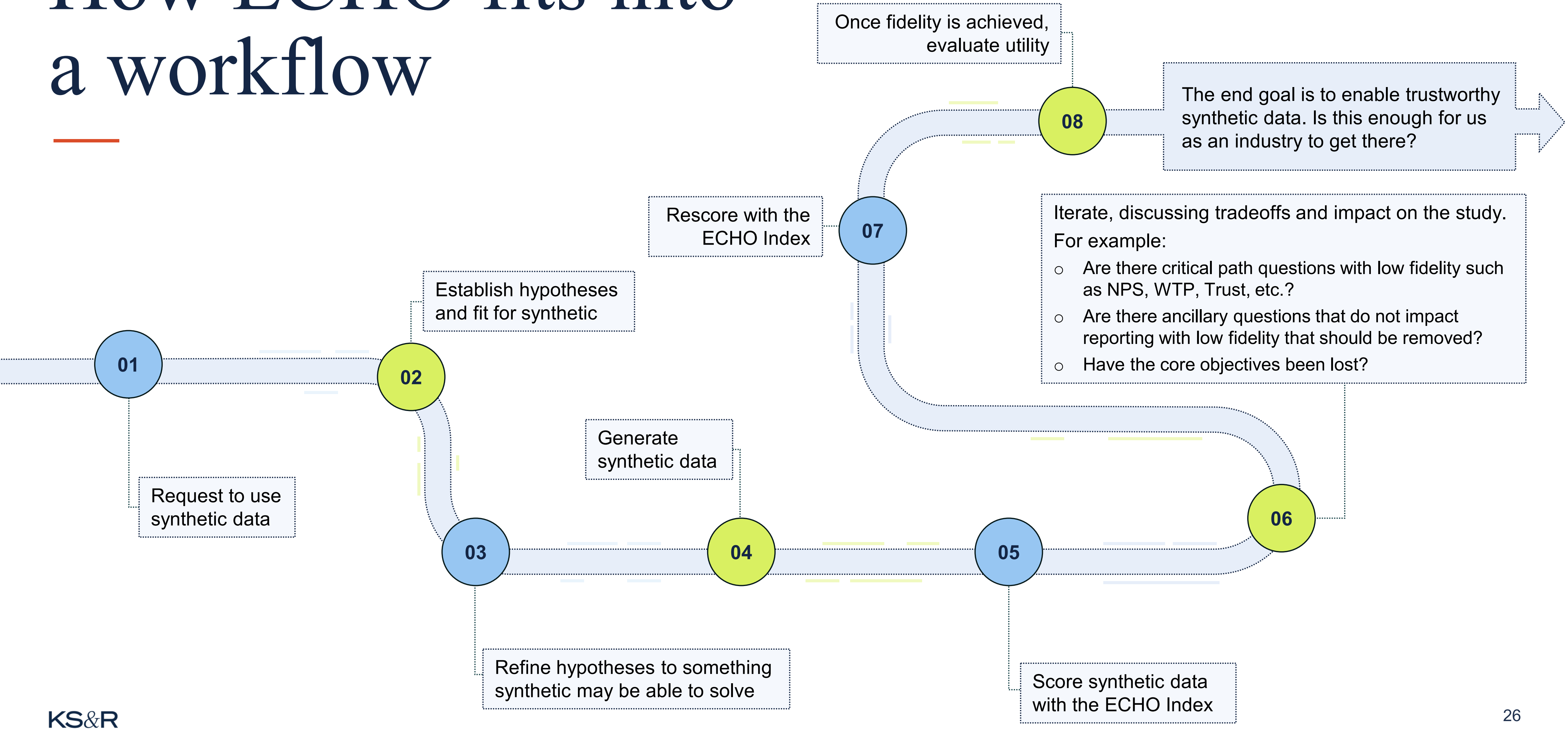
All synthetic provider models showed high fidelity for Paxton and Yang when evaluated with the ECHO Index.

Comprehensive, complex data

It is difficult to produce high fidelity data across the board, especially for complex and nuanced lines of questioning. A review of the strengths and weaknesses of a synthetic model is critical to build trust in synthetic data.

The Bottom Line: Governance is a requirement to understand when synthetic works.

How ECHO fits into a workflow



Experimenting with the ECHO Index has surfaced several best practices.

What works

- **The model and provider matters.** There is no clear criteria to determine when deep-learning / ML approaches or LLMs will be most successful.
- **Question type matters.** Likert / agreement scales align well; take caution with other questions such as single-selects or demographics.
- **No clean base-size rule.** Sub-segment boosting looks promising, but there are many factors to consider.
- **Expect iteration.** Like segmentation, good synthetic work takes tuning and collaboration.

Pitfalls to avoid

- **Scope too broad.** Rescope into bite-size, testable hypotheses first, and be prepared to refine on the fly.
- **No real-world training data.** It's a requirement, not a suggestion.
- **Highly complex studies.** Long surveys, nested logic, advanced analytics. Focus on core insights and KPIs.
- **Vendor self-assessing.** Use an independent governance framework to verify their claims.

The Bottom Line: Solve the “easy” problem well before reaching for synthetic insights.

What’s Coming Up Next

Sample boosting was just the starting point. Here's what else we're working on.

- The State of Synthetic Data full report — named-provider performance — in September
- A new study aimed at striking a balance between the two evaluated datasets to further explore synthetic data fidelity
- Extension of the ECHO Index to apply to digital twin models
- Full methods in the SRC conference proceedings this fall

We're keeping the evaluation open. If you use or build synthetic data, put your models to the test with us.

Interested in more?

Thinking of scoring your own data?

Reach out and we'll show you how your synthetic models perform against real data.

Ben Cortese, PhD

Decision Sciences & Innovation
Senior Vice President
bcortese@ksrinc.com